BENGALI ACCENT RECOGNITION FROM SPEECH

BY

S.M. Saiful Islam Badhon
ID: 162-15-7878

Md. Habibur Rahaman
ID: 162-15-7761

Farea Rehnuma Rupon
ID: 162-15-7707

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Sheikh Abujar
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised By

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

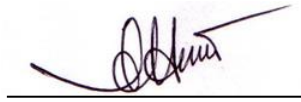
DHAKA, BANGLADESH

JULY 2020

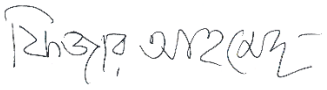
APPROVAL

This Project/internship titled **"Bengali Accent Recognition from Speech"**, submitted by S.M. Saiful Islam Badhon, Md. Habibur Rahaman, Farea Rehnuma Rupon, ID No: 162-15-7878, 162-15-7761, 162-15-7707 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 09 July 2020.

BOARD OF EXAMINERS



Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Fizar Ahmed — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Tarek Habib — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

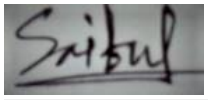
DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Sheikh Abujar, Lecturer (Senior Scale),** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

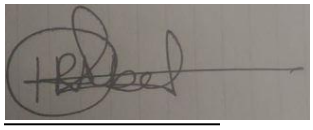
Supervised by:

Sheikh Abujar
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

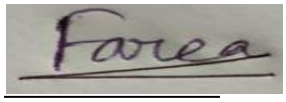
Submitted by:



S.M. Saiful Islam Badhon
ID: 162-15-7878
Department of CSE
Daffodil International University



Md. Habibur Rahaman
ID: 162-15-7761
Department of CSE
Daffodil International University



Farea Rehnuma Rupon
ID: 162-15-7707
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Sheikh Abujar, Lecturer (Senior Scale),** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of machine learning, natural language processing and data mining to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Syed Akhter Hossain, Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Communication between human and computer is the most vital term of current world and the performance of computer or machine mostly depends on this communication. Undoubtedly, human speech is the most comfortable form of communication for human and that's why the world is now trying to reduce the dependencies on text by using speech communication which producing a huge amount of audio data. Human speech is nothing, but analog signals, different wavelength of this signal represents different age's speaker, different gender's speaker even different language's word has different wavelength. And the question was "dose different accent of same language produces different wavelengths"? we tried to find out the answer of this specific question. For that we used different accents of Bengali language. The services of voice based applications are very limited in Bengali language for maximizing the of benefit of voice based application in Bengali we need to train our machine with local Bengali language and for that we need to perfectly identify that which accent actually speaker is speaking. But the problem is these digital machines don't handle the analog signals. That's why we need to convert the signal into numeric value for that we used a very popular and effective techniques of feature extraction which is Mel Frequency Cepstral Coefficient and for classification we used different classification algorithms. We got maximum 86% accuracy on 9303 data of different classes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

Research has been interested identifying and synthesizing accents for decades but that automatic speech recognition system (ASR) was developed until the middle of the twentieth century. In natural language processing (NLP) ASR system are used for identifying speech. NLP is a section of artificial intelligence (AI) which plays a vital role in involving computing techniques and linguistics to interpret natural languages such as humans. The NLP term has been extended to several areas of science and technical advancement, such as text classification, classification of images, and voice identification etc. [1].

Identification of accent is one of the widely discussed topics in NLP. A language's accent represents the inhabitants of a regional area and a social or economic class to where they include. By incorporating a variety of strategies from other aspects of speech technology, different methods to automated accent detection were investigated. The motive behind previous work has been developed automated voice identification systems, as significant degrees of accent variance in each language may be tough for a system to cope with [2].

Among the world, Bengali language is in the fifth position as the most spoken local language and in the seventh position for the most spoken language. Approximately 228 million people talk in the local Bengali language and more than 37 million people use Bangla as their second language [3]. There are several forms of endemic accents in Bengali language. Dhaka (old), Khulna, Chittagong, Mymensingh, Sylhet, Barisal, Rangpur and Rajshahi are the leading spoken dialects in almost all of the Bangladesh.

This research establishes an application using the regional Bengali dialects except Rangpur and added standard Bangla to identify a person’s region using the wave frequency of a person’s speech. For feature extractions MFCC has been used here and different algorithms were applied to get the best accuracy for the system.

1.2 Motivation

In this era, there are so many facilities that can be got by human speech. Speech identification is one of the most important sectors of NLP. Several works have been done with voice recognition. Some AI assistant like Siri [23], Alexa [5] etc. have been developed but these are not applicable for detecting different accents or dialects from voice. But with Bengali language there are only few works and the accuracy of those systems are very poor.

As Bangla is one of the well-known languages in this world and it has a wonderful history behind this language there should be a good voice recognition system for this language. In Bangladesh there are many regional dialects. Most of the AI assistants which supports Bangla, those are made only for standard Bengali language. So rural people can't get support from those systems who are not able to speak the standard Bangla properly.

So, considering this situation we decided to construct a system that can be able to detect the Bengali regional accents from the wavelength while they talk. The objective of this research is to build an effective system with a well-off dataset that can build a human-computer interaction so that the rural people can be benefitted and gives the best accuracy.

1.3 Rationale of the Study

Many researches have been concluded on Natural Language Processing but most of them are in English language. These methods or the mechanisms are being used in several automatic systems. There are almost 300 million people around the world who speak in Bangla but lamentably a very few researchers work on it and most of them have a poor accuracy. Moreover, there are several dialects that vary to the areas of Bangladesh. People from different regions speak different accents. There is a great lack of Bengali accent detection systems in NLP.

Voice-based applications are playing a significant role in recent period. The applications may be helpful in various ways such as AI assistant, surveys for census, in finance, HR and marketing, medical sector, public transportation, auto typing, solving crimes etc. To maximize the uses of voice-based applications using Bengali dialects for the welfare of the rural people we get interested to work with this.

1.4 Research Questions

- When people express the same language in various accents do the vocal folds emit various wavelengths?
- How can we collect the data of different dialects?
- Which methods can be applied for the feature extractions?
- Which algorithm should be used for classification to find the best result?
- Can this study be effective for the rural inhabitants of Bangladesh?

1.5 Expected Outcome

The prospective result of this research-based project is to develop an application for the general population of Bangladesh to know which region they belong using the wavelength of their vocal cord. This type of application may help in many ways. Such as:

- This application may help in collecting demographic statistical data such as population of that area, educational background, gender etc.
- Several kinds of misdemeanors happen via telephone calls and audio clips which can be categorized by their accent and endorsed under trial.
- This research can be beneficial for the tourist centers and foreign automatic telecommunication services as they can use it for finding the region of a person by identifying the accent.

1.6 Report Layout

In **chapter 1**, this report discusses about what we are going to do, why we are going to do and how we are going to do. Overall, the motivation behind this work with expected outcome is described briefly in this chapter.

In **chapter 2,** related work of this sector has been described. And summarizing their work findings from their works also noted in this section. By finding their limitation we set our goals by explain the challenges.

In **chapter 3,** this report discusses about methodology has been used in this work. Some theoretical topics are also discussed in this chapter which are related to this work. Process of data collection, data preprocessing, feature extraction, methodologies used in this work are briefly discussed in this chapter.

In **chapter 4,** the result came from previous chapter have been presented and comparison and best process also showed in this chapter.

In **chapter 5,** summery of the project is main focus. Future work, conclusion, limitation and recommendation also noted in this last chapter of the report.

CHAPTER 2

Background

This chapter will describe related works with our project. There are many researches has been completed about accent classification from various languages. We already know that various language has various accent. Depending on those accents many researches has been done. We will describe the summary of those research in this chapter. Additionally, the project model, strength, weakness of those research will be discussed as well.

2.1 Related Works

Identifying accent from language is not a new research anymore. Many researchers are already tried to identify accent in many ways. In this paper we used many machine learning and deep learning algorithm to figure out the best possible accuracy. Mainly we used Linear Regression, Decision Tree, Gradient Boosting, Random Forest and Neural Network for the classification.

In last few years many good researches have been done by different researchers around the world. R. K. Mamun et al. worked on Bangla Speaker Accent Detection by MFCC [4]. They mainly collected many speeches from different accent of Bangladesh. They have collected voice from Old Dhaka, Barishal, Noakhali, Chattagram, Rajshahi, Sylhet, Mymensingh, Khulna, Rangpur which was 16 kHz sample rateand 16-bit quantization. All those voices were collected in a noise free environment and those sample was mainly some responses that was asked by the volunteers. Figure 2.1 showed the main block diagram of MFCC method.

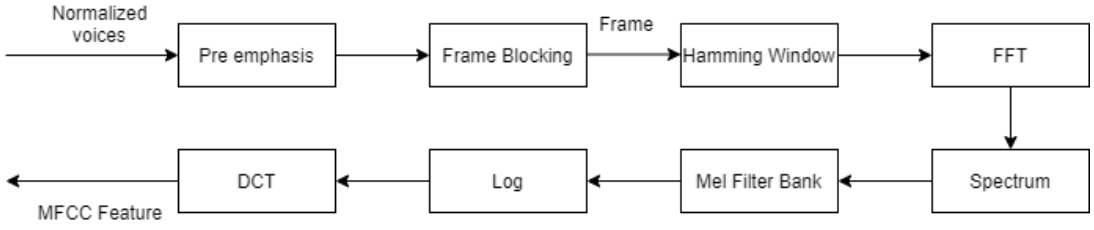


Fig. 2.1 MFCC Block Diagram [4]

Then they have performed RNN to differentiate individual dialects after applied the MFCC. At this method it's mainly checked voice activity at several times and this process was continued till it cannot finds similarity of voices. Figure 2.2 shows method of their dialect recognition system.

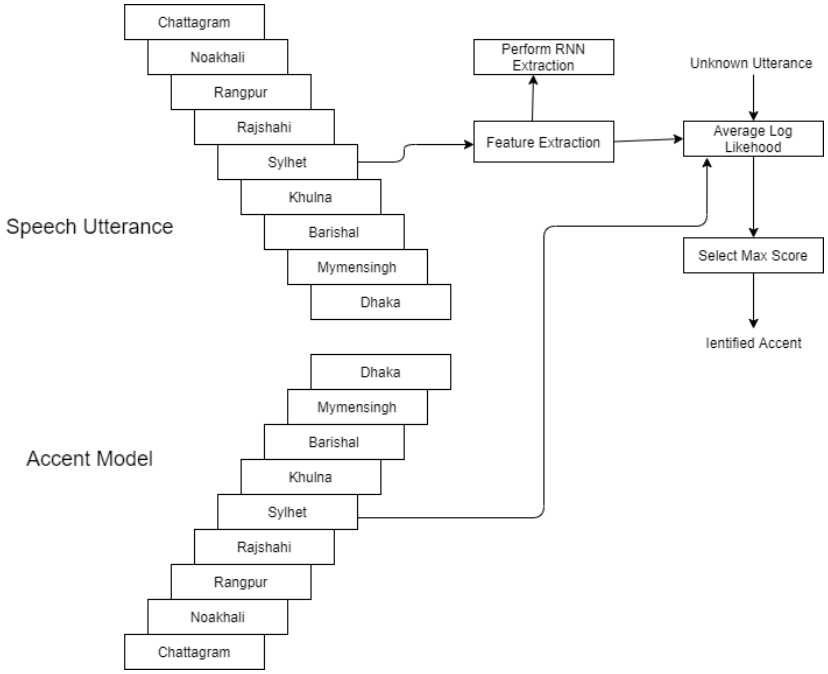


Fig. 2.2 Dialect Recognition System [4].

They have got some false alert rate in their system. They mentioned minimum experimental result if the accuracy rate is higher. On the other hand, they mentioned average result when accuracy rate and error rate (EER) is equal. Table 2.1 illustrated the results of the detection percentage of the accent variation.

TABLE 2.1: EXPERIMENTAL RESULT [4]

Dialects	Accuracy (%)	EER (%)
Chattagram	69	31
Sylhet	75	25
Rajshahi	77	23

Khulna	68	32
Barishal	83	17
Mymensingh	72	28
Rangpur	76	24
Noakhali	78	22
Dhaka	72	28

Another researcher named Georgina Brown [2], examines total five different automatic accent recognition system. It distinguished between geographically proximate accents. By geographically proximate accents is expected to increase the degree of similarity between a type of task which may be used to forensic speech analysts. This was the first main part of this research work. The second and last part is mainly concerned with identifying the specific phonemes which are important in a given accent recognition task and eliminating those which are not. Mainly they classified the phonemes which are most useful to the task depending on the varieties. Among of those five systems the first system is GMM-UBM system. This system trained with multiple speaker's speech including all accents involved. MFCC was used for the feature extraction. The data was extracted from 25ms windows of speech at overlapping 10ms intervals (Figure 2.3). Second system is GMM-SVM. This system is almost the same system that I've mentioned in the first system just the enrolment data is speaker-specific but still independent of speaker content. Rather than adjusting one single model to rep-loathe one highlight, a model is adjusted for every one of the speakers in the enrolment information (Figure 2.4).

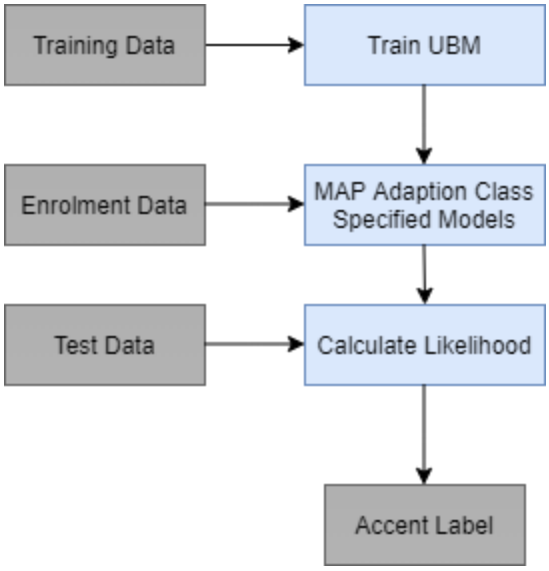


Fig 2.3: flow diagram of GMM-UBM [2]

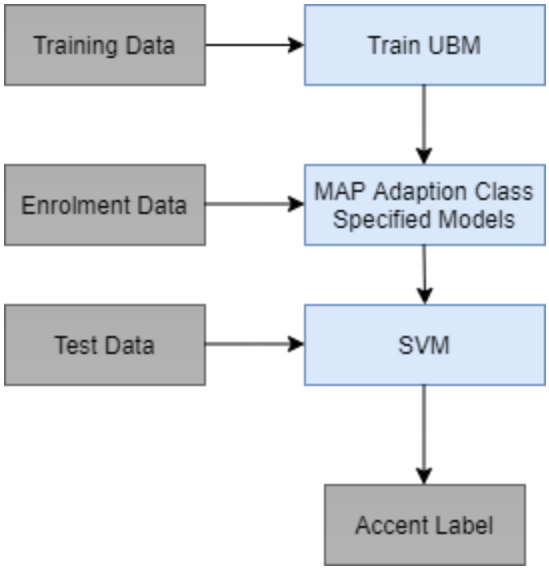


Fig 2.4: flow diagram of GMM-SVM [2].

In the third system, the preparation speakers' discourse tests, alongside their ortho-realistic interpretations, for every one of the accents included are taken and constrained adjusted. Utilizing these arrangements, a GMM is prepared to speak to every phoneme for an individual speaker. All the GMM implies for every phoneme are connected to speak to the speaker's articulation framework in one long super vector. Fourth system is an orthographic translation and discourse test for every speaker in an emphasize class is gone through a constrained aligner. The midpoint 12-component MFCC vector for every vowel telephone is extricated and a normal midpoint MFCC vector is determined for every English vowel phoneme. Last and the fifth system- Speakers are handled as in the fourth system to show a delegate ACCDIST framework for every speaker. The contrast between systems 4 and 5, notwithstanding, lies in the arrangement procedure. For each highlight class, the speaker networks having a place with that class are taken care of into an SVM (similarly as the GMM-SVM framework) and the ACCDIST grids for every single other speaker of every single other emphasize are taken care of in to shape a 'one-against-the rest setup. Finally, the experiments result is in the table number 2.2.

TABLE 2.2: RECOGNITION RATES GENERATED BY ACCENT RECOGNITION SYSTEMS IN PAST STUDIES.

System	Accuracy (%)	No. Classes
GMM-UBM	61.13	14
GMM-SVM	76.11	14
Phon-GMM-SVM	63.2	5
ACCDIST-based-Corr	93.17	14
ACCDIST-based-SVM	95.18	14

On account of the Phonological-GMM-SVM framework from, the possibility desire joined to the outcome above is20.0%, as their investigation included recognizing just five Flemish assortments. Given this, the after effect of 63.2% doesn't appear as noteworthy. The standard order rate right now 80.8% right, appearing differently in relation to the Y-ACCDIST-SVM result introduced above, where just vowels were utilized (86.7% right). The flat line is the gauge acknowledgment level of this complement acknowledgment task when all vowels and consonants are handled through the framework (80.8% right). SVM-RFE appears to all the more reliably accomplish an acknowledgment rate above gauge, while ANOVA accomplishes the most noteworthy acknowledgment rate in general (89.2% right).

TABLE 2.3: RECOGNITION RATED FOR EACH ACCENT RECOGNITION SYSTEM CLASSIFYING THE 40 TEST AISEB SPEAKERS INTO ONE OF FOUR ACCENT GROUPS (25% CORRECT EXPECTED AT CHANCE).

System	%Correct
GMM0UBM	37.5
GMM- SVM	35.0
Phon-GMM-SVM	62.5
Y- ACCDIST -Corr	82.5
Y- ACCDIST -SVM	87.5

Another researcher Ayodeji Olalekan Salau, Tilewa David Olowoyoand Solomon Oluwole Akinola [6], talks about the means that were taken in building up the highlight arrangement strategy utilizing a one-dimensional convolutional neural system (1D CNN) and long momentary memory (LSTM) organize model (1D CNN LSTM).

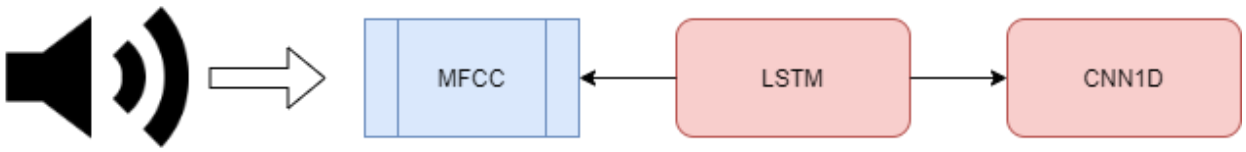


Fig 2.5: Overview of the 1D CNN method

Sound information as discourse were gathered from 100 speakers from every one of the three significant Nigerian indigenous dialects, in particular: Hausa, Igbo, and Yoruba. Boldness programming was introduced in the chronicle gadget. Spared all the procured discourse information into an organizer which we named "gained discourse information," the information were brought into the Audacity programming to evacuate foundation clamor and from that point onward, the discourse information was sent out from the mp3 position into the .wav group.

TABLE 2.4: THE MODEL PERFORMANCE FOR THE THREE DIFFERENT ACCENTS.

Model approach (1D CNN LSTM)	Hausa accent (%)	Igbo accent (%)	Yoruba accent (%)	Average (%)
Training Accuracy	98	91.5	97.2	95.6
Test Accuracy	97.7	90.6	96.4	94.9

TABLE 2.5: RESULTS OF VALUES OBTAINED FROM FEATURES.

Features	Value
Data Format	.wav

Mode	Mono
Audio Rate	20,000
N_MFCC	40

The proposed 1D CNN LSTM arrange model with the planned calculation had the option to perform an order of speakers into Hausa, Igbo, and Yoruba giving a normal exactness of 97.7%, 90.6%, and 96.4%, separately.

2.2 Comparative Analysis and Summary

So, it's clear that a lot of research has been done about accent classification. Many researchers have done it in many ways. In most of the cases, MFCC was used for feature extraction. And then many researchers have given many shapes to the data. Some of them collect data from the noise-free environment and some of the researchers collected data from a normal environment and then they have given a new shape to those data. At the same time, some researchers used software like Audacity to process the data and some researchers use many algorithms or processes to give data in a new shape. Though in every language have many accents but it's really tough to work with all of them. In Bangla Language, it also has so many accents. Some of the researchers have used those accents by division wise and some of them just research on many divisions. At the same time, it's also clear that collecting data is not that much easy. Because there are hugely lacking data resources. At last, we also noticed that most of the data were used for training purposes, and only a few data used for testing purposes. And thus, all the researchers have reached a successful result. Figure 2.6 is showing the summarized genaral process of all researchers.

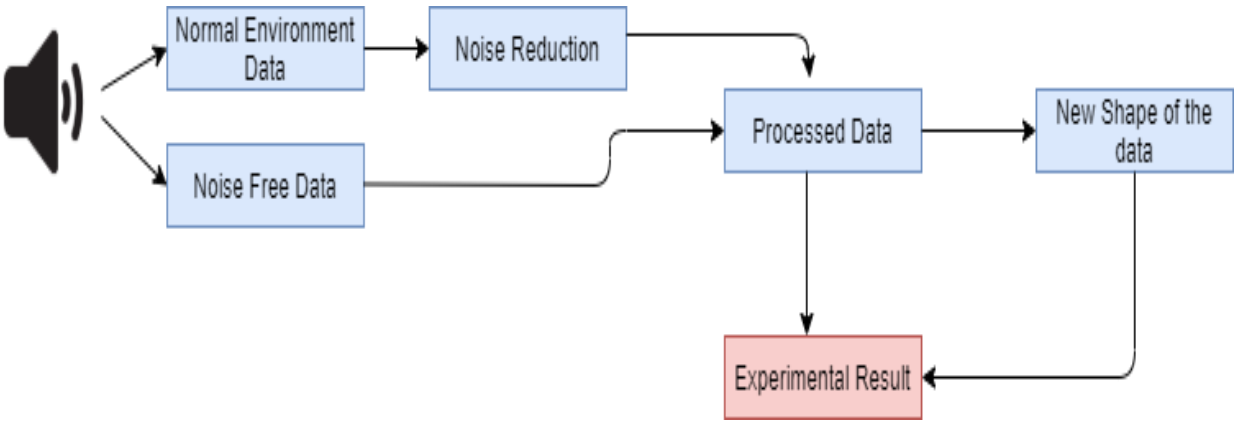


Figure 2.6: In general process

2.3 Scope of the Problem

Accent detection may help in a lot of ways. In some cases, it may help to identify criminals or identify the most dangerous zone for the crime. Like, if we got any criminal then through this accent classification system, we could detect the zone from which the criminal belongs. And are some cases, this can be the one best clue to reach the success level of a crime. On the other side, this system can help with marketing. Sometimes businessmen have only some targeted people for their business and for this business through this system anyone can identify the region of his targeted customer. Through this, he/she can make a profit by business. So definitely this system helps to identify the targeted people for the businessman.

2.4 Challenges

1) **Data Collection:** The most common challenges are collecting the data. In Bangla Language, data collection is the main challenge. Because there are no organized datasets for Bangla Language. At the same time, the source of collecting the data for different accents is not sufficient enough. For collecting the data, we had to go to many different places. Because it's not possible to collect all the different accents at the same time in the same place. Additionally, for this project we needed some special data like music free voice, noise-free voice, etc. and this was also a big

challenge because we cannot collect any kind of data. So, we had to take help from many YouTube videos, Google, etc.

2) **Collect the exact accent:** As we mentioned before, for this project we had to need a special kind of data and this was also the main challenge. Other hands, nowadays people mix up different accents at the time of utterance. That's why we had to reject those voices/speeches.
3) **Making Language Compatible with system:** Bangla grammar is much more complex than English. So, we had to face a lot of problems to shape the data in a proper manner and then we have faced problems to compatible the data with the system. But the good thing is, we researched a lot for this and then we learned a lot.
4) **Model Selection:** Though a lot of research has been done about accent classification it's still challenging research for Bangla Language. Because most of the research has done in English Language and that's why a lot of models has been already introduced. It was a difficult task to select the proper model for Bangla Language which will provide the best rate of accuracy.
5) **Proposed Workflow:** As we had to find out the best workflow for this project, so we had to experiment in many processes. The workflow of this project is given below:

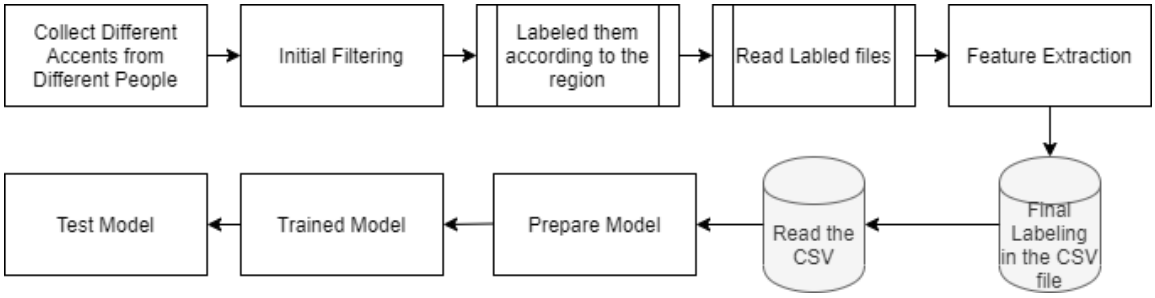


Figure 2.7: Workflow of the project

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

In this point we will discuss about the instruments and techniques we used for collecting our research data. Data is the most important element of any type of research and when it's about Natural language processing using machine learning it requires huge amount of data. We collected total 9303 voice of different classes (Dhaka, Khulna, Barisal, Rajshahi, Shylet, Chottogram, Mymensingh, Noakhali, Formal-bangla). It was tough to collect human voice manually that's why follow three different techniques to collect the voices from speakers.

1) **Manually:** First approach was very straight forward we collected voice from speaker manually with scripts. But the problem was it was too time consuming and its quite tough to find out our required accent.
2) **Google Form:** As in our first approach it was really tough to find out the require accents, we start collecting data through google forms from different area speakers. But in both techniques, it was really time consuming and speakers were not willing to produce audio for long time.
3) **Youtube video:** For better and enough data we start collecting data from youtube videos [7] where we got different accent and different type of speakers for that you used a website yt-cutter.
4) **Normalization:** we need to convert all the audio file into same format and same sample rate so that we can work on same type of data. All audio files were converted in wav format and 16000 Hz sample rate. For this work we used a software called audacity [8].
5) **Segmentation:** And finally, we segmented the audio files into 5 seconds. So that we can compare same type audio files. For segmentation we used a procedure which is given in figure 3.1.

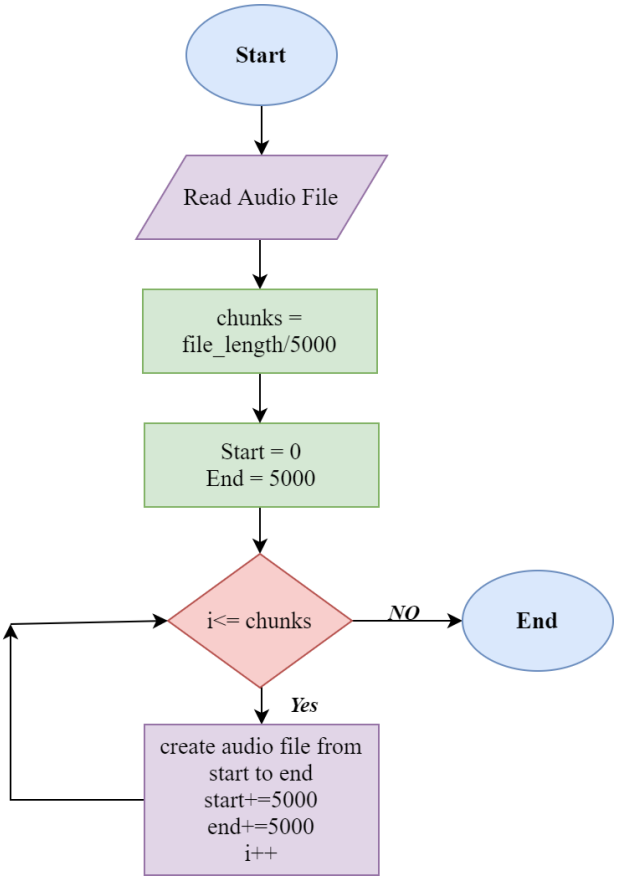


Fig 3.1: Segmentation procedure

3.2 Dataset Utilized

In this experiment we used audio data. One of the most critical form of data to process for machine learning. We had total 9303 data of different classes and the duration was 4-5 seconds. Age range of speaker was 20-60 years old. The most important part of audio dataset is extracting features from audio files. For feature extraction we used MFCCS. We extracted 26 features from every audio file the features were chroma feature, Root Mean Square Error, spectral centroid, spectral bandwidth, roll off, zero crossing rate and 20 features from MFCCS.

1) **Chroma feature:** This feature helps to find out the cords from audio signals [9]. By calculating cords this feature can identify thickness of voices. Depending on

thickness we may classify the accents of different area's people. Normally, longer wavelength produces thicker pitches which our ear identifies as thick voice. Comparison of figure 3.2 will give us an idea about this.

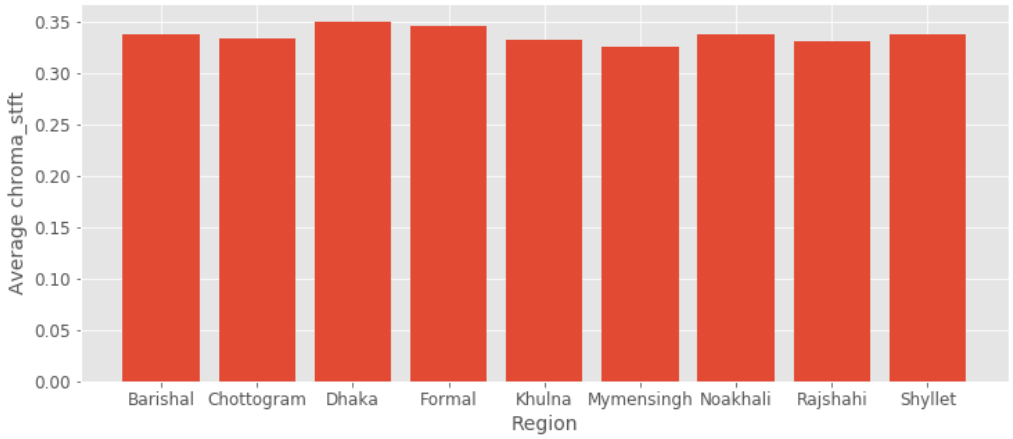


Fig 3.2: Comparison of chroma feature of different accents

Above figure is comparing the average values of chroma feature of different region. That showing the difference is not so high.

2) **Root Mean Square Error:** This is not an audio feature but very useful values for any type of data science researches. It finds out difference between actual value and observed values [10]. By calculating these values, it can measure which values are close to each other which may help us to classify features into different classes. For calculating this we used below formula:

$$RMSE_{fo} = \sum_{n=0}^{n} [(z_f - z_o)^2/N]^{1/2} \qquad (1)$$

Here, $(z_f - z_o)^2$ = Difference square and N = sample size.

3) **Spectral centroid:** This feature extracts the center of mass of an audio file. That helps to find out the loudness of a speech [11]. As we are working with language,

language hasn't any loudness issue, but the speaker of different area has their own way of speaking some area's people speak louder than normal some speak less loudly than needed. This was really a very important feature for this classification problem. Comparison of Spectral centroid given in figure 3.3.

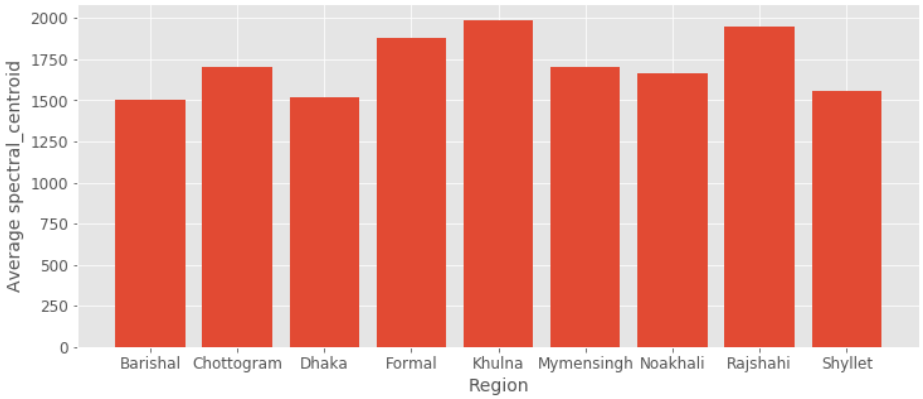


Fig 3.3: comparison of spectral centroid of different accents

4) **Spectral bandwidth:** The difference between lower and higher point of an audio signal is called spectral bandwidth [12]. Below (figure 3.4) boxplot representation will give us an idea difference between different area's voice in respect of spectral bandwidth.

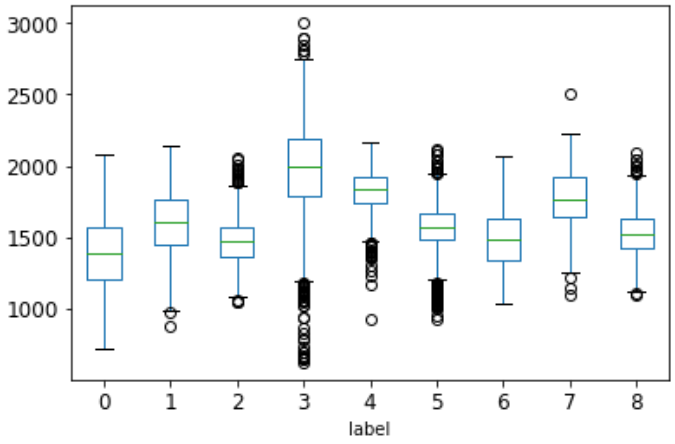


Fig 3.4: Box-plot representation of spectral bandwidth

5) **Roll off:** Spectral roll off filters frequency energies in specified points. It is normally 85% [13]. Figure 3.5 will describe difference in roll-off of different accents.

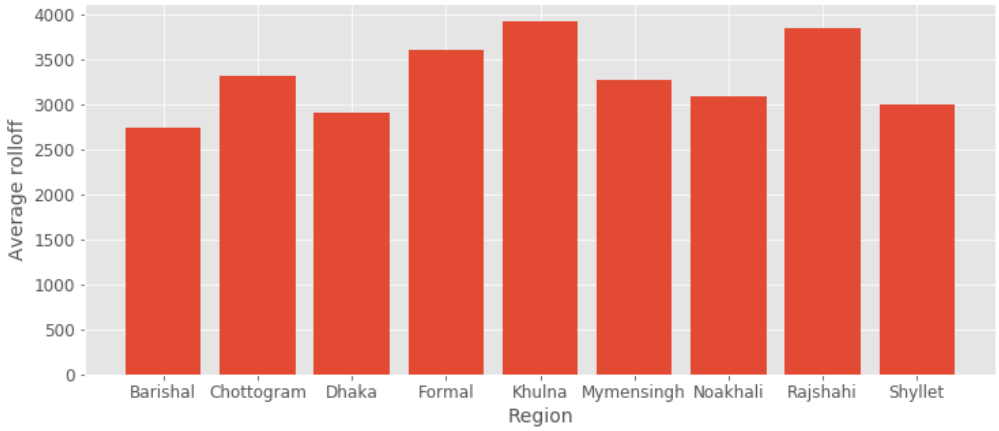


Fig 3.5: comparison of spectral roll-off of different accents

6) **Zero crossing rate:** This feature finds out the changes of signs in an audio signal [14]. Broadness of an audio signal can be finding out by this feature. This may help classifying deferent type of speaker. For accent classification it's not a proper feature to use but we took as much as possible features for better results.

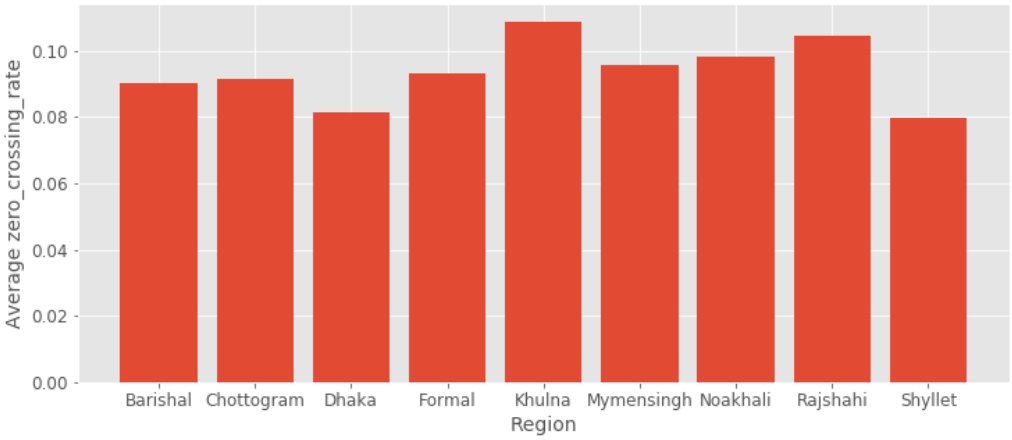


Fig 3.6: comparison of zero crossing rate of different accents

Figure 3.6 also showing very less difference between ZCR of different accents.

7) **MFCCS:** This is the most important feature. This actually not a single feature but it's a packet of 20 individual features [15]. We need to follow a procedure (figure 3.7) for finding the mfccs features.

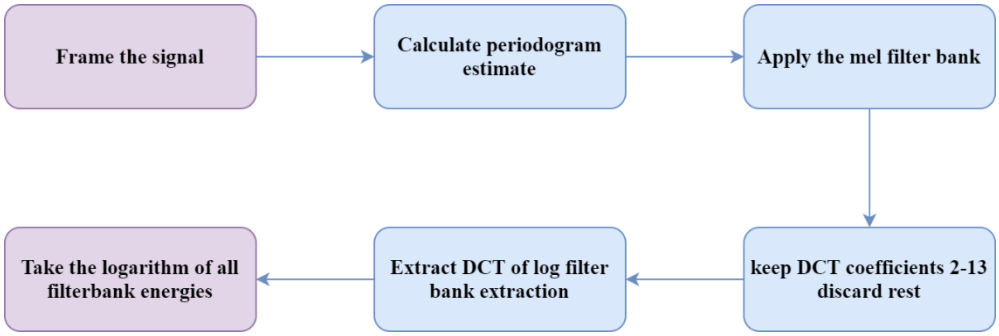


Fig 3.7: Procedure of MFCCs extraction

For starting the procedure, we need to convert the normal frequency in mel-scale. The equations are given below.

$$M(f) = 1125 \times \ln(1 + f/700) \quad \text{.... (2)}$$

$$M^{-1}(m) = 700 \times \left(exp(m/1125 - 1)\right) \quad \text{.... (3)}$$

Equation (2) is for converting normal frequency to mel-scale and equation (3) is for mel-scale to normal frequency.

The steps of extracting MFCCs are given below,

- At starting we need to segment the whole audio signal in small frames. Standard size of frames is 25 MS but 20 to 40ms also acceptable. So, if we have 16kHz signal we will get 16000*0.025 = 400 frames.
- Now we need to apply discreate fourier transform (DFT) to the frames. For that we have to follow an equation.

$$Si(k) = \sum_{N}^{n=1}\left(Si(n)h(n)e^{i2\sqcap kn/N}\right) \quad \text{…. (4)}$$

In equation 4 we find Si(k) where, I = number of frames, Pi(k) = power spectrum. Si(n) = time domain frame by frame, K = frame length. For extracting power spectrum using Si(k) we need below equation.

$$Pi(k) = \frac{1}{N} \times |S_i(k)| \quad \text{…. (5)}$$

- This step will find out the filter-bank which is basically set of 20-40 filters. And filter bank is power spectrum which we found in previous steps.
- Log filter bank energies will be the log value of filter bank that we find in previous step.
- And finally, by converting those final values into discrete cosine we will find final objective t cepstral coefficients.

3.3 Statistical Analysis

This section will discuss about dataset's detail information. We wrote about amount of data and classes of our working dataset previously. Table 3.1 is containing some more information about the dataset.

TABLE 3.1: DATASET INFORMATION

Region	Represented as Number of	Total Speaker	Speaker male	Speaker female	Total Data
Barishal	0	67	29	38	1257
Chottogram	1	39	25	14	797
Dhaka	2	26	10	16	763
Formal	3	89	49	40	1652
Khulna	4	52	32	20	750
Mymensingh	5	42	26	16	1053

Noakhali	6	55	20	35	1213
Rajshahi	7	21	13	8	860
Shylet	8	27	14	13	958

Bar chart visualization is given bellow for better understanding.

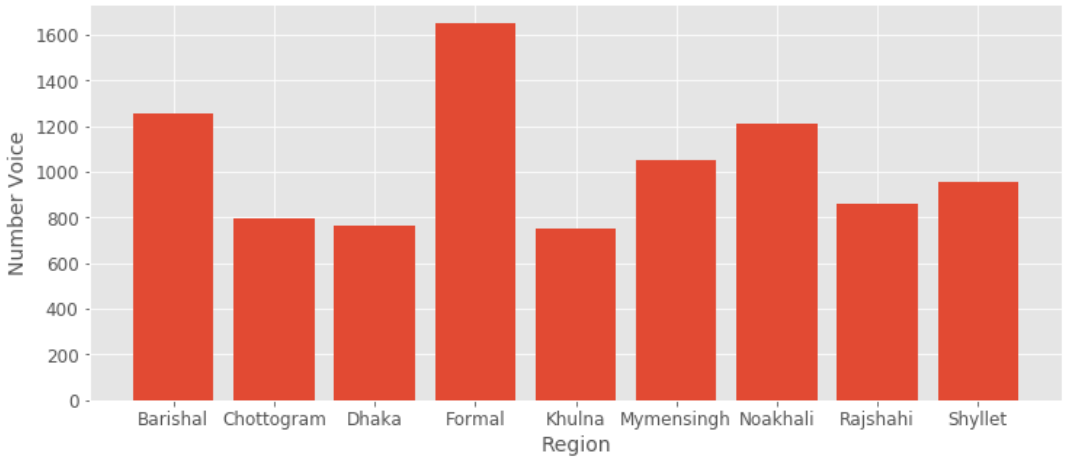


Fig 3.8: Region wise number of data

In figure 3.8 it's clear that we collected lots of Formal accent and less Khulna accent and we missed Rangpur language due to less amount of speaker and resources. Below pie chart (figure 3.9) will show ratio of male and female speakers.

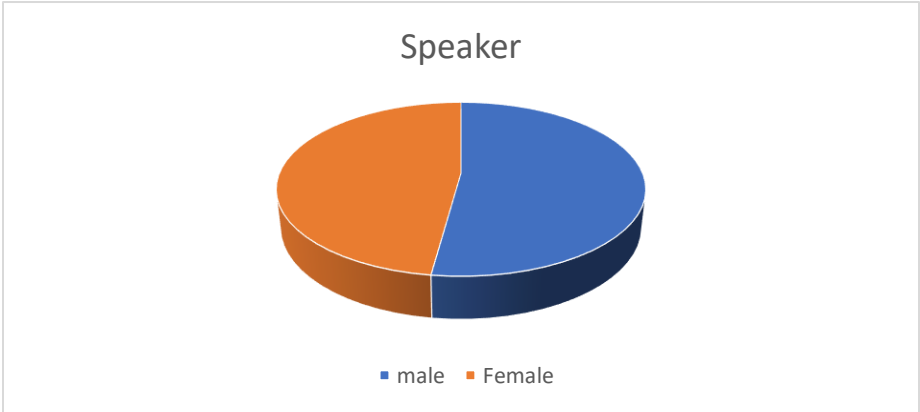


Fig 3.9: Ratio of male and female speaker in dataset

3.4 Proposed Methodology

This work followed a procedure or methodology for getting ultimate result. Figure 3.10 will discuss about our methodology briefly.

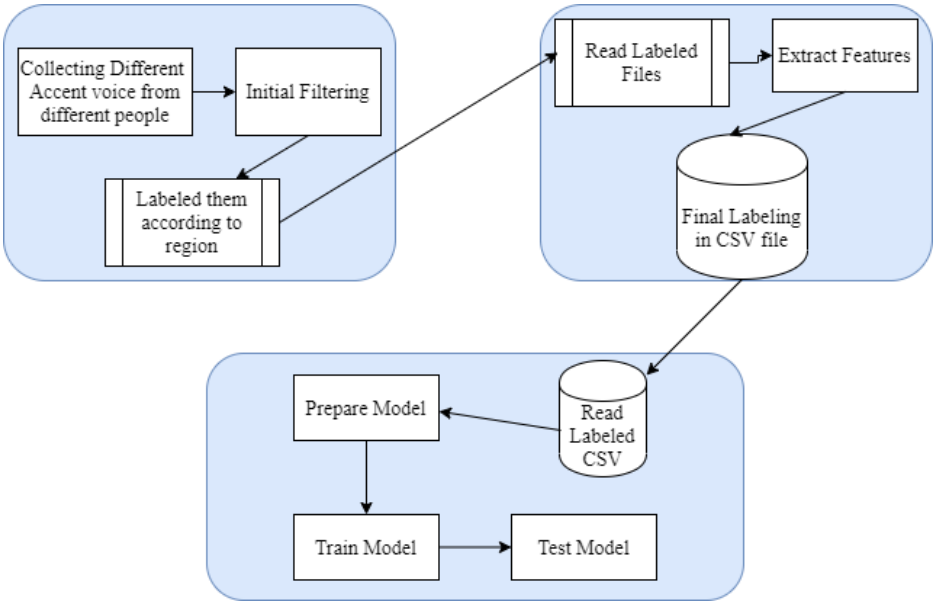


Fig 3.10: Proposed method

According to the figure, our first work is collecting raw data and label them by storing into folders and then the most important and complex work which is extracting the features and label them according to region name. And at last train model using those data.

3.5 Implementation Requirements

Python 3.8:

Python is a high-level programming language [16]. It can be used for desktop GUI and web applications but most importantly it has a rich resource in data science and machine learning. Which actually helps us to complete our work easily. The less complexity and easiness of python programming language increase its acceptance to all the tech enthusiast. And we used 3.8+ version in our research work which is the updated version of the time.

Anconda 4.5.12:

This the free and open source distribution of python [17]. This is also available for R programming language. This actually a bundle installer. By installing single thing its install lots of necessary tools for data science. Even it comes with a concept of virtual environment. We can isolate different projects from each other so that we can use different requirements for each of them. We used 4.5.12 version of anaconda, the updated version of the time.

Jupyter notebook:

We used jupyter notebook for writing the code [18]. This actually a web based open source which allows to write codes, visualize the data, using equations and lot more. We used 6.0.3 version of jupyter notebook.

Keras:

keras is a deep learning library which written in python language [19]. This library actually makes the works easy. They already developed the calculation parts we just use them in proper place according to our necessary. In backend, keras is using tensorflow. There is some other library whose are using keras as backend but keras is most developed and rich. We used keras 2.3 with tensorflow 2.0 version. This work used keras and tensorflow for experimenting the deep neural network on its dataset.

Scikit learn:

This is a python free library which features various classification, clustering and regression algorithm [20]. It makes easy to use those algorithms. We used multiple algorithms on our work from the library. The used algorithms are listed below:

- Support vector machine
- Logistic regression
- K-nearest neighbor
- Random forest

- Gradient boosting

Pydub:

Pydub is a free python library [21]. Normally used for simple audio processing. There are some functionalities of pydub are listed below:

- Read audio files
- Play audio files
- Split audio files
- Marge audio files and lot more.

Librosa:

Librosa is a feature extraction library from signals [22]. Computers can't work with analog data, so we need to convert them into numeric values. Librosa did this work for us. We extracted 26 features from a single audio file. We used 0.7.2 version of librosa.

Audacity:

This a free and open source audio editing software [8]. We used it for converting the audio format of the files. For merging the discreate files and most importantly for normalizing by converting in 16kHz. We used 2.3.3 version of audacity.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

In previous chapter we completed our data collection, preprocessing, feature extraction and set a methodology to reach the goal. According to the procedure now we need to apply different algorithms we tried Random forest, gradient boosting, K neighbor classifier, Logistic regression and neural network. Random forest, gradient boosting and neural network shows better result so we will discuss about these three procedures in this section.

We tried random forest in our dataset. We used 80% of data for training purpose and 20% for testing the model. Bellow figure 4.1 will give us an idea about this.

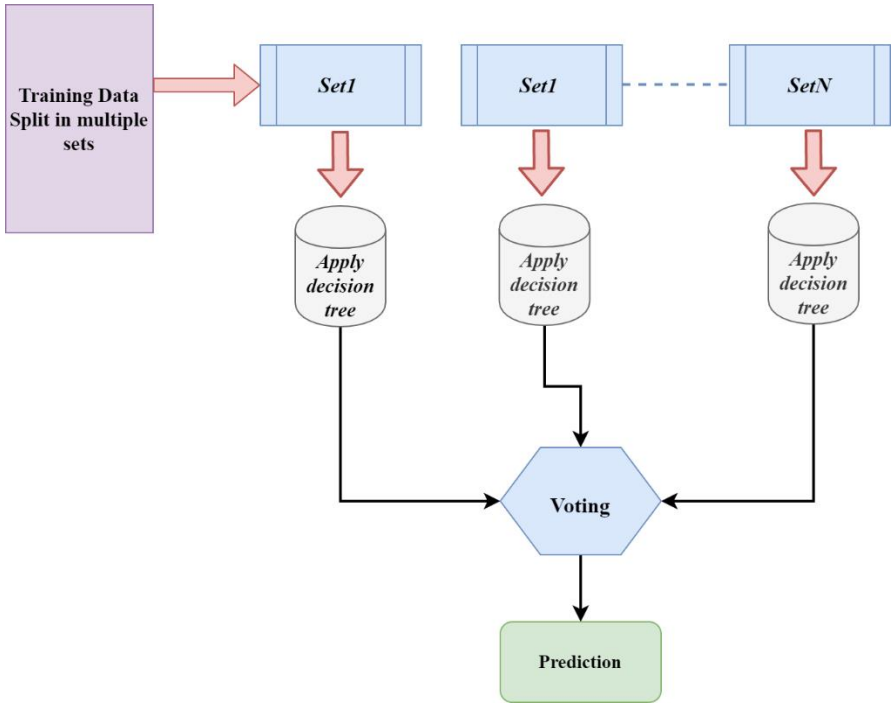


Fig 4.1: Random forest procedure

According to the above figure, dataset was divided into multiple set of data and decision tree was applied to them individually and after completing all there was another decision taking phase where we found the best result.

After testing on random forest model with 20% data of the dataset we got below(figure 4.2) confusion matrix which will give us an idea about the performance of the model.

Fig 4.2: Confusion matrix of Random forest

Gradient boosting is another classifier we tried. This is actually an updated and modified version of previously developed Ada-boosting. This is nothing but decision tree, this algorithm actually tries to improve the performance of decision tree by multiple iterations. Figure 4.3 of confusion matrix will give us better understanding of it.

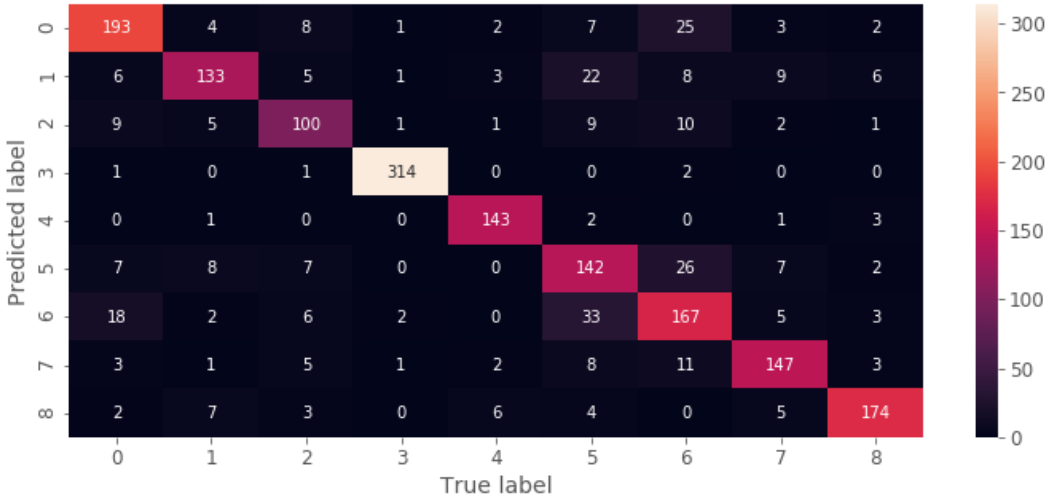


Fig 4.3: Confusion matrix of Gradient Boosting

Figure 4.3 is showing less accurate result than our decision tree, so our experiment went wrong at that point. But till then it was clear that random forest was given us better result.

Then we tried deep neural network. We used three hidden layers with input and output layers. Below figure 4.4 will describe the model information properly.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_26 (Dense)             (None, 32)                864
_________________________________________________________________
dense_27 (Dense)             (None, 16)                528
_________________________________________________________________
dense_28 (Dense)             (None, 16)                272
_________________________________________________________________
dense_29 (Dense)             (None, 15)                255
_________________________________________________________________
dense_30 (Dense)             (None, 9)                 144
=================================================================
Total params: 2,063
Trainable params: 2,063
Non-trainable params: 0
```

Fig 4.4: Deep Learning model summery

And below figures (4.5 and 4.6) will show us the accuracy and loss of the model according to epochs. We used 500 epochs for this model.

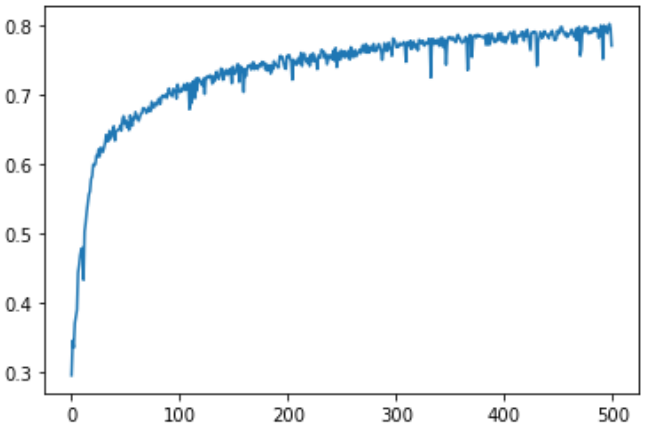


Fig 4.5: Deep learning model accuracy graph

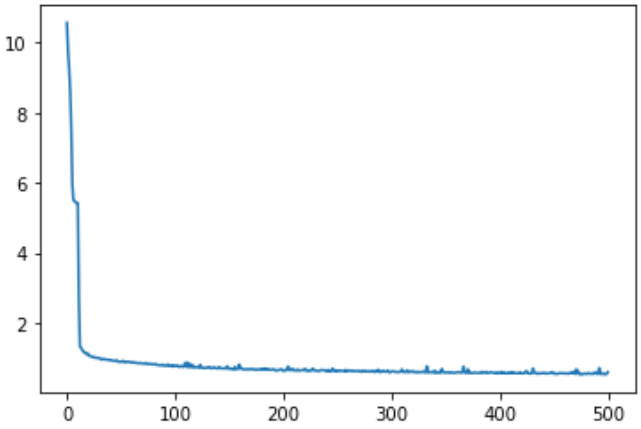


Fig 4.6: Deep learning model loss graph

And finally, confusion matrix of figure 4.7 will give us a better understanding about the performance of our DNN model.

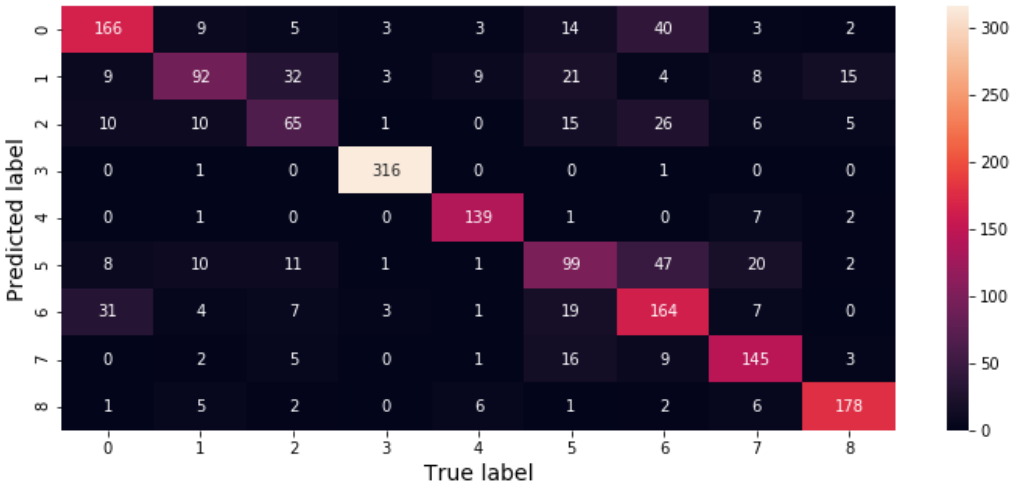


Fig 4.7: Confusion matrix of deep learning model

4.2 Experimental Results & Analysis

We applied multiple algorithms on our dataset and random forest is showing better result. In table 4.1 we compare the classifiers accuracy.

TABLE 4.1: COMPARISON OF DIFFERENT CLASSIFIERS

Algorithm	**Precession**	**Recall**	**F1-score**	**Accuracy**
SVM	0.37	0.41	0.38	0.45
Logistic regression	0.47	0.48	0.47	0.52
KNN	0.57	0.57	0.56	0.58
Random forest	0.86	0.85	0.85	0.86
Gradient boosting	0.80	0.80	0.81	0.81
Neural Network	0.81	0.81	0.81	0.81

Classification report of random forest will give better view of understanding which is given in below figure 4.8.

	precision	recall	f1-score	support
0	0.83	0.83	0.83	245
1	0.94	0.75	0.84	193
2	0.82	0.74	0.78	138
3	0.98	0.99	0.99	318
4	0.92	0.98	0.95	150
5	0.70	0.76	0.73	199
6	0.73	0.78	0.75	236
7	0.86	0.88	0.87	181
8	0.93	0.92	0.93	201
accuracy			0.86	1861
macro avg	0.86	0.85	0.85	1861
weighted avg	0.86	0.86	0.86	1861

Fig 4.8: Classification report of Random forest model

CHAPTER 5

Summary, Conclusion, Recommendation and Implication for Future Research

5.1 Summery of the Study

The research project is developed for predicting accent of different area of Bangladesh. For maximizing the use of ASR system in all level of Bangladesh it's important to identify the accents of speakers beside that there are lots of application of this. We used MFCCs techniques for extracting the features from speeches and used multiple machine learning and deep learning techniques for feeding the machine. We split our dataset into 80% and 20% ratio for training and testing. This work used 80% for training and 20% for testing the system.

5.2 Conclusion

Bangla is one of the most spoken language in the world [3]. In Bangladesh bangla is the first language and some part of India also speak in bangla [4]. Native speakers have their own way of speaking. This way or style is different according to the areas. Huge number of speakers produce huge number of audio data and data is valuable when its processed. We just tried to find another feature of audio data by this work. If we look at previous work of bangla in accent classification there is very small amount of data with average accuracy of 74.44% in this work we have made an improvement in data set (total 9303) which helped us to improve the accuracy which is 86%. And we added formal bangla language which is missing in previous work.

5. 3 Implication for Further Study

We must try make things better always. Even in this work there are lots of things can be improved. Some future work is listed below:

- Increasing the number of accents. In bangla there are lots of accent we worked with very few of them.
- We need more verities in speaker. Verity give a model more strength.
- Trying more classifier and comparing them.

REFERENCES

[1] Salau A.O., Olowoyo T.D., Akinola S.O. (2020) Accent Classification of the Three Major Nigerian Indigenous Languages Using 1D CNN LSTM Network Model. In: Jain S., Sood M., Paul S. (eds) Advances in Computational Intelligence Techniques. Algorithms for Intelligent Systems. Springer, Singapore.

[2] Georgina Brown (2016), Automatic Accent Recognition Systems and the Effects of Data on Performance, Odyssey 2016, June 21-24, 2016, Bilbao, Spain

[3] "Bengali language, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Bengali_language [Last accessed: 1- May- 2020]

[4] Mamun R.K., Abujar S., Islam R., Badruzzaman K.B.M., Hasan M. (2020) BanglaSpeaker Accent Variation Detection by MFCC Using Recurrent Neural NetworkAlgorithm: A Distinct Approach. In: Saini H., Sayal R., Buyya R., Aliseri G. (eds)Innovations in Computer Science and Engineering. Lecture Notes in Networks andSystems, vol 103. Springer, Singapore

[5] "Alexa" [Online]. Available: https://www.alexa.com/ [Last accessed: 1- May- 2020]

[6] Salau A.O., Olowoyo T.D., Akinola S.O. (2020) Accent Classification of the Three Major Nigerian Indigenous Languages Using 1D CNN LSTM Network Model. In: Jain S., Sood M., Paul S. (eds) Advances in Computational Intelligence Techniques. Algorithms for Intelligent Systems. Springer, Singapore

[7] “Youtube” [Online]. Available: https://www.youtube.com/ [Last accessed: 1- May- 2020]

[8] “Audacity” [Online]. Available: https://www.audacityteam.org/ [Last accessed: 1- May- 2020]

[9] “Chroma feature” [Online]. Available: https://en.wikipedia.org/wiki/Chroma_feature [Last accessed: 1- May- 2020]

[10] “RMSE: Root Mean Square Error” [Online]. Available: https://www.statisticshowto.com/rmse/ [Last accessed: 1- May- 2020]

[11] “Spectrul centroid” [Online]. Available: https://librosa.github.io/librosa/generated/librosa.feature.spectral_centroid.html [Last accessed: 1- May- 2020]

[12] “Spectral Bandwidth” [Online]. Available: https://www.sciencedirect.com/topics/engineering/spectral-bandwidth [Last accessed: 1- May- 2020]

[13] “Spectral Features” [Online]. Available: https://musicinformationretrieval.com/spectral_features.html [Last accessed: 1- May- 2020]

[14] “Zero crossing” [Online]. Available: https://en.wikipedia.org/wiki/Zero_crossing [Last accessed: 1- May- 2020]

[15] “Mel Frequency Cepstral Coefficient (MFCC) tutorial” [Online]. Available: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ [Last accessed: 1- May- 2020]

[16] “Python” [Online]. Available: https://www.python.org/ [Last accessed: 1- May- 2020]

[17] “Anaconda” [Online]. Available: https://www.anaconda.com/ [Last accessed: 1- May- 2020]

[18] “Jupyter notebook” [Online]. Available: https://jupyter.org/ [Last accessed: 1- May- 2020]

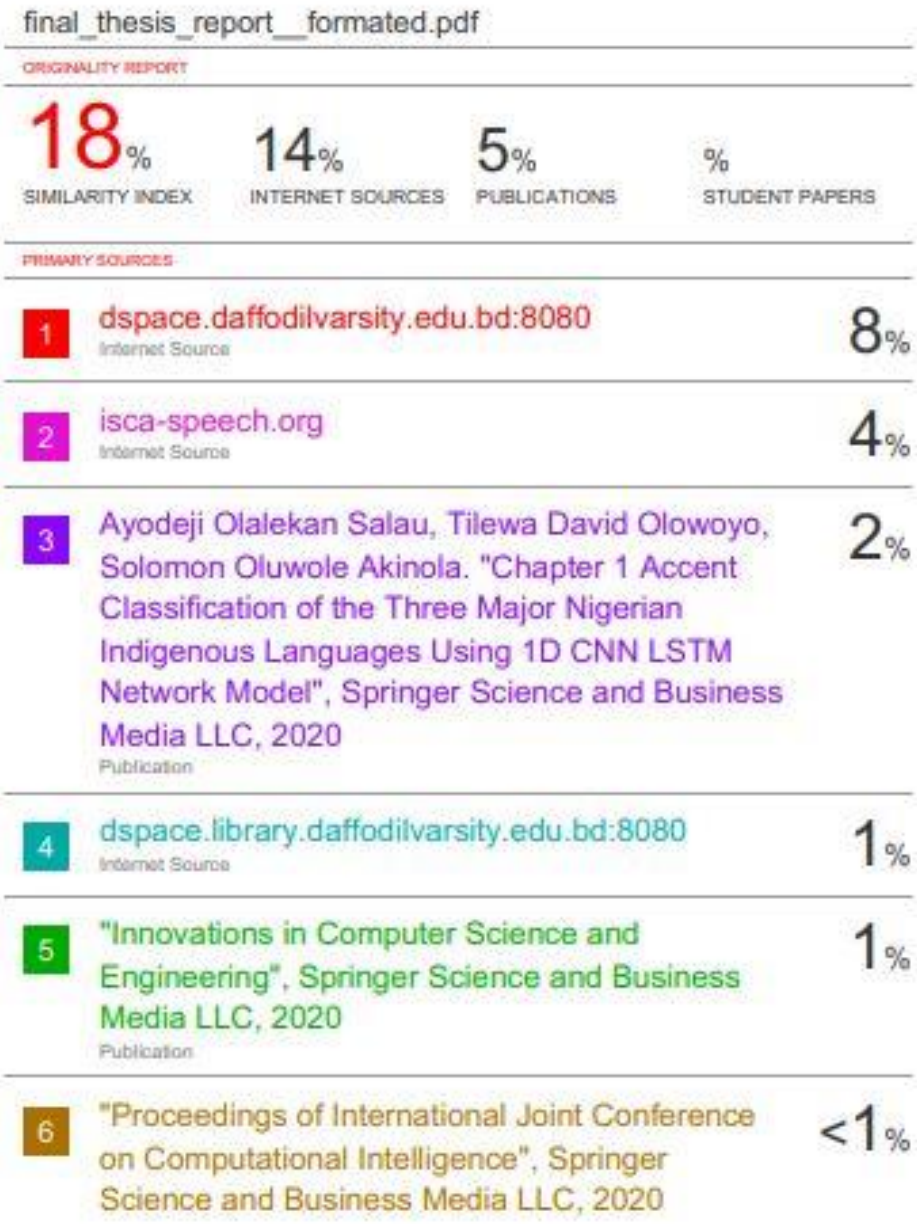
[19] “keras” [Online]. Available: https://www.tensorflow.org/guide/keras [Last accessed: 1- May- 2020]

[20] “Scikit learn” [Online]. Available: https://scikit-learn.org/stable/ [Last accessed: 1- May- 2020]

[21] “Pydub” [Online]. Available: https://pydub.com/ [Last accessed: 1- May- 2020]

[22] “Librosa” [Online]. Available: https://librosa.github.io/librosa/ [Last accessed: 1- May- 2020]

[23] “Siri” [Online]. Available: https://www.apple.com/siri/ [Last accessed: 1- May- 2020]



final_thesis_report__formated.pdf

ORIGINALITY REPORT

18%	14%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	isca-speech.org Internet Source	4%
3	Ayodeji Olalekan Salau, Tilewa David Olowoyo, Solomon Oluwole Akinola. "Chapter 1 Accent Classification of the Three Major Nigerian Indigenous Languages Using 1D CNN LSTM Network Model", Springer Science and Business Media LLC, 2020 Publication	2%
4	dspace.library.daffodilvarsity.edu.bd:8080 Internet Source	1%
5	"Innovations in Computer Science and Engineering", Springer Science and Business Media LLC, 2020 Publication	1%
6	"Proceedings of International Joint Conference on Computational Intelligence", Springer Science and Business Media LLC, 2020	<1%