Speech Emotion Recognition using Librosa Library

BY

Mijanur Rahman Polin
ID: 191-15-2685
AND

Md. Mashrafi Hassan
ID: 191-15-2705

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

Supervised By

Mushfiqur Rahman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

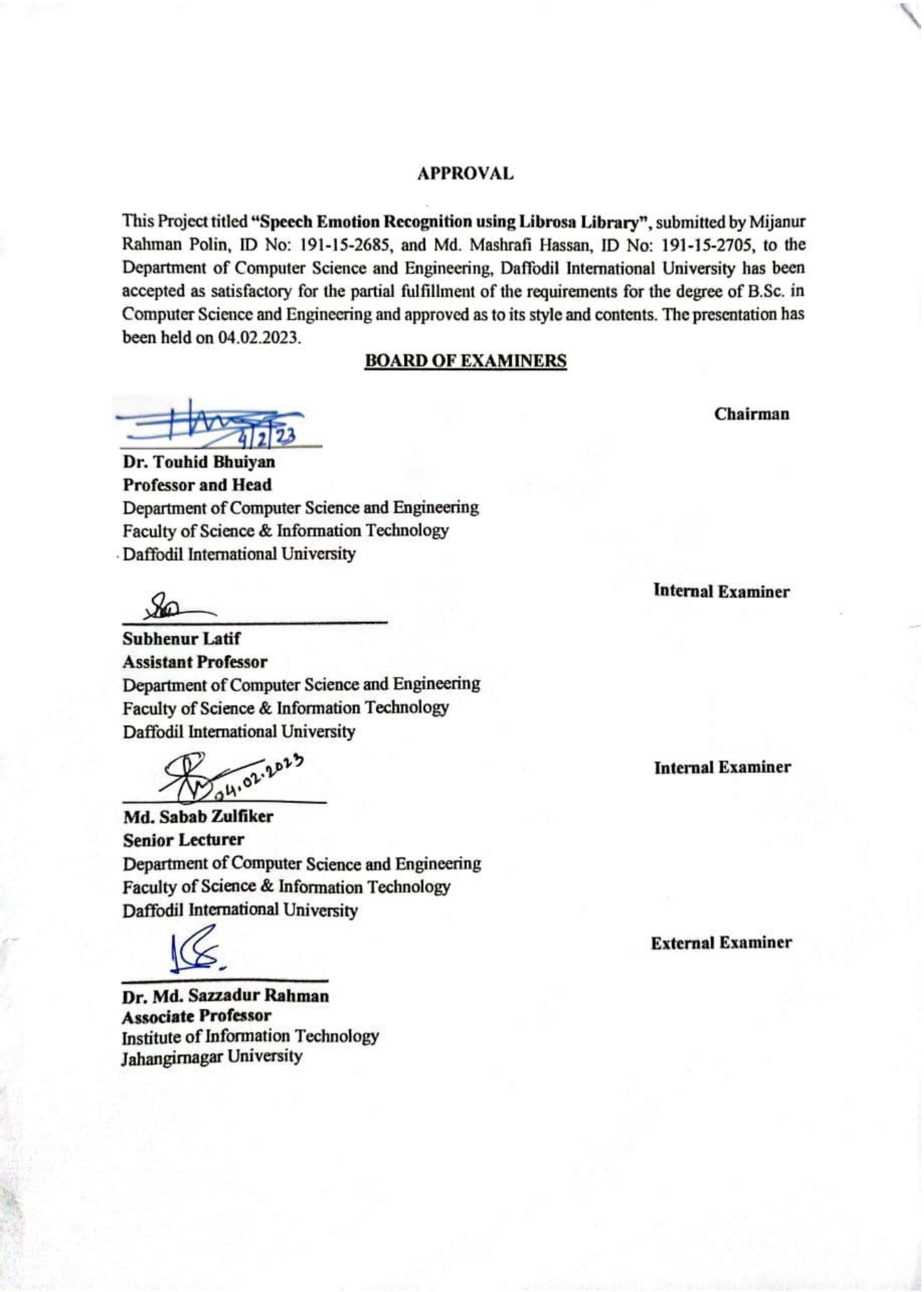
Co-Supervised By

Md. Mahfujur Rahman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023



APPROVAL

This Project titled **"Speech Emotion Recognition using Librosa Library"**, submitted by Mijanur Rahman Polin, ID No: 191-15-2685, and Md. Mashrafi Hassan, ID No: 191-15-2705, 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 04.02.2023.

BOARD OF EXAMINERS

4/2/23

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Subhenur Latif
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

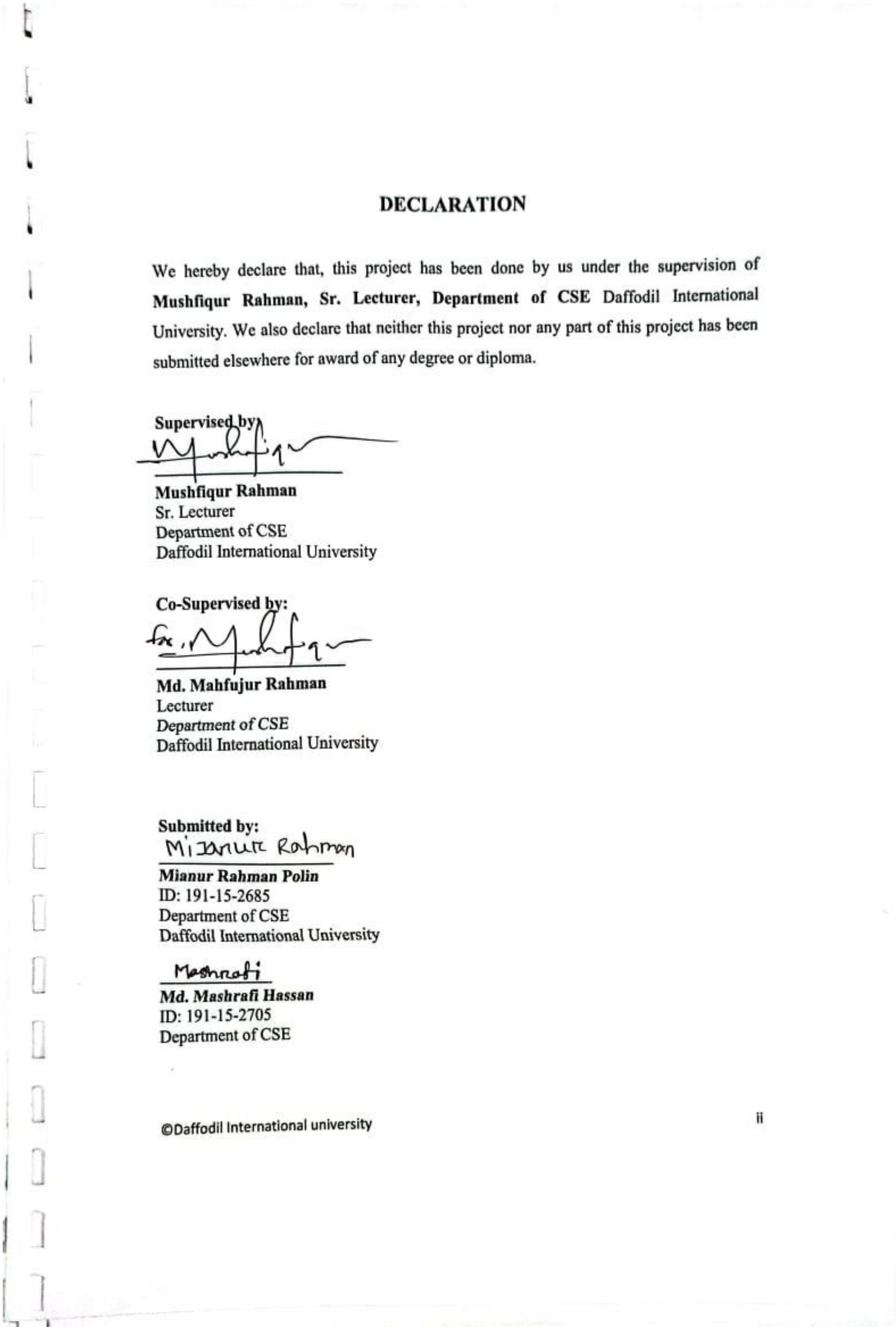
04.02.2023

Internal Examiner

Md. Sabab Zulfiker
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md. Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University



DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mushfiqur Rahman, Sr. Lecturer, Department of CSE** 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:

Mushfiqur Rahman
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Mahfujur Rahman
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Mianur Rahman

Mianur Rahman Polin
ID: 191-15-2685
Department of CSE
Daffodil International University

Mashrafi

Md. Mashrafi Hassan
ID: 191-15-2705
Department of CSE

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 **Mushfiqur Rahman, Sr. Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "**Deep Learning**" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Mushfiqur Rahman, Md. Mahfujur Rahman, and 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

All living human being on earth express their opinion through a language. Some meaningful voice makes a language. In all the call center giving service faces many problem when they talk to customer. Sometimes the employee do not understand the emotion that the customer express. From the past two decades speech emotion recognition from speech becomes an interesting topics to researcher’s. This study is about various deep learning based algorithms RELU and SOFTMAX as well as models called Long-Short Term Model (LSTM) for the purpose of recognizing the emotion from speech. The dataset (TESS) is consist of 5600 voice data which is divided in seven categories such as (fear, angry, disgust, neutral, sad, pleasant-surprise, happy). The LSTM gives accuracy of 99.20% and it is the highest accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Speech Emotion Recognition (SER) is a task that can recognize the emotion from speech. Naturally human uses their speech for expressing their opinion. Speech is one of the best things in the world to communicate with each other. Human use it efficiently for the purpose of communication. We need a device that is programmable which is automatically recognize the voice. A programmable device must need for the purpose of research. Millions of people use mobile phone to communicate with each other. While communicating with each other, people uses various kinds of voice containing emotion like happy, sad, angry, bore, pleasure, interest, wonder. Car drivers, pilots, naval pilot in sea, astronauts in the spaceship uses their voice to communicate with each other. In call center operations and customers also express their communication. If we ever noticed, call center employees never talk in the same manner, their way of talking to the customers changes customers. This also happened to common people as well as. The employees recognize customers emotion from speech, so they can improve their service and convert more people. Thousands of known and unknown emotions are creating in the day to day life. In the medical sector doctor, nurses, technicians try to identify the emotion of the patient and give proper treatment to the patient. We know now a days online gaming becomes very popular to the young generation. For the improvement of gaming experience we need to recognize the voice and find out the emotion. In stock market, stress monitoring and much more applications arounds us. Social media also becomes popular day by day. In social media like facebook, twitter, Instagram and other platform we use voice chat, video chat or conversation which contains voice. This project is designed to recognize the emotions in people’s verbal speech. We have been using cutting-edge technologies to develop an AI system that can accurately identify the emotion being expressed in a person’s speech. This project will help us to better understand our interactions with others and to provide better support for those in need. The project uses deep learning to analyze the audio signals from a person’s speech. We use several features extracted from the audio to train a machine learning model to predict the emotion being expressed. We also use natural language processing to understand the context of the speech and to better identify the emotion. We hope that this project will help us to better understand the emotions of others and to provide better support for those in need.

1.2 Motivation

Emotions are very initial for human being, in our day to day life we use it such as communication with each other, learning something new as well as for decision making. The emotions are expressed through our voice, facial expression, gestures and other clues. There is a set of objectively measurable parameters in voice which reflect the affective case of a person is currently expressing. For example, happy often produces the sign in face as well as make the body more flexible, changes the voice as well as vocal tract shape. Previously, emotions were measured by human being physically and is was not research by computer scientists. Now a days the field has been update. It creates a new field for the new researcher. These steps create a new era for the people who are suffering with autism. Now we can detect angry caller or who uses bad language. A new method of emotion recognition is proposed in this paper. To find the emotion by measuring and analyzing the voice is very hard work. In the past the research of recognizing the emotions entirely subjective. In communication, special attention is required to find the clues the speaker conveys. Our target is to create efficient, real time methods that can recognize the emotion of the mobile phone users, car drivers, pilots, specially call center customers as well as many other human machine communications.

1.3 Rationale of the Study

We have already studied many papers related to speech emotion recognition (SER). Each of the research paper discuss different task like voice extraction, spectrogram of the training and test data. Some paper also discusses on another topics. For getting the best accuracy by applying different algorithms, we are motivated by these papers and get the best accuracy in our Speech Emotion Recognition (SER). Our goal is to detect emotion that helps in office specially call center. The study of speech emotion recognition is important because it can provide insight into how people perceive the emotions of others. By understanding the nuances of speech and how it can be used to convey emotion, researchers can better understand how people interact with each other. Additionally, this research could be used to create more effective communication systems that could improve the effectiveness of computer-mediated conversations. Finally, this research could help to improve the accuracy of automated systems that rely on speech recognition technology.

1.4 Research Questions

- What are the most effective methods for recognizing emotions from speech?
- How can output be improved?
- What are the advantages and disadvantages of speech emotion recognition?
- What benefits have you got since you started using emotion recognition?

1.5 Expected Output

To recognize the emotion from speech which can detect various kinds of emotion that helps in offices like customer call center and they will improve their customer service. We hope that our system will accurately extract the speech and find the best accuracy.

1.6 Project Management and Finance

Project management and finance are two related fields of study. Project management focuses on the planning, organizing, and controlling of projects, while finance is focused on the management of money and other assets. Both skills are important for successful business operations. Project management involves the use of project planning tools, risk management strategies, and quality control methods. Financial management involves budgeting, forecasting, and decision-making. Both fields require the use of data to make informed decisions and create effective strategies. Project management and finance work together to ensure that resources are used efficiently and that projects are completed on time and within budget.

1.7 Report Layout

- Chapter 1 covered the main ideas of "Speech Emotion Recognition using Librosa Library,". Besides about our study's purpose, goal, and outcomes.
- In chapter2 section, the brief summary, problem, and outcomes are discussed from previous related research works.
- The research methodology is covered in Chapter 3.
- The details of the experimental findings are detailed in Chapter 4.
- Chapter 5 describes our social impact on environmental effects, sustainability.

- Chapter 6 describes Summary, Conclusion, Recommendation, and Implication for Future in this research.
- Reference

CHAPTER 2

Background

2.1 Preliminaries

Speech emotion recognition (SER) is a technology used to identify the emotion in a person's speech. It uses algorithms to analyze the speech and determine its emotional content. The technology can be used to provide insights on human emotions in a variety of contexts, such as customer service, healthcare, education, and marketing. It can also be used to detect emotional states in real-time, allowing for automated systems to respond to the emotional state of the speaker. Many researchers work on different topics of deep learning. Nowadays deep learning, artificial intelligence are known as recognized technologies. For recognizing speech emotion, deep learning method is best for getting highest accuracy. Gather a large dataset of speech recordings with emotional labels. Preprocess the speech recordings using signal processing techniques, such as feature extraction. Develop and train a model using a machine learning algorithm to recognize speech emotions. Test the model's performance using cross-validation or another method of evaluation. Refine the model and evaluate its performance on new sets of data. Deploy the model in an application or system.

2.2 Related Works

Authors of [1] had used deep learning techniques to extract emotions from signals that includes many well known speech analysis as well as classification techniques. They used deep rural network, deep learning, deep Boltzmann machine, recurrent neural network, deep belief network, convolutional neural network. These methods elaborated emotions like happiness, joy, sad, neutral, disgust, surprise, anger, fear. They have some limitations like internal architecture, less efficiency for temporally verifying input data as well as over learning when memorize the information.

Authors of [2] proposed a system that addressing three valuable aspects , firstly suitable features for the speech recognition, secondly design an appropriate classification and last one was proper preparation that includes emotional speech database. They had some limitations which were the uses of vectors that were statistically independent. This intension was bad practice. They could improve the classification performance by using autoregressive models.

Authors of [3] developed automation analysis which could recognize human affective behavior. They used a number of systems, algorithms and models. They use SVM, SFS and MFCC.

Authors of [4] proposed Hidden Markov models. They proposed two methods, first one was global statistics which was classified by Gaussian mixture models second one was increased temporal complexity. They only use 6 emotion as a result when we use other speech emotion it could not recognize. They might improve their research by using all other emotion.

Authors of [5] had worked on speech signals to find out frequency, energy, duration of silence and voice quality. They also proposed a method that makes short time long frequency power coefficients (LFCC) to represent the speech signals and a discreet Hidden Markov model (HMM) as the classifier.

Authors of [6] had said that static or dynamic classification problem could be removed by using speech emotion recognition. They also describe a frame based formulation that relies minimal speech processing. They should use convolutional neural network for getting better performance.

Authors of [7] proposed a deep learning system (CNN) to get hierarchical representation. They had faced some distortion for the reason of environment. They got the highest accuracy by using (CNN) method.

Authors of [8] prepared Fourier parameter model which was used determine the first and second order differences. They improve the accuracy rates by using by using Mel-frequency Cepstral Co-efficient (MFCC). They used EMODB, CASIA, EESDB database for their model.

[9] They proposed real time speech emotion recognition in human to computer interaction. They had some limitation like datasets can not be successful like real world scenario.

Authors of [10] used Deep neural network (DNN) to extract high level features from raw data. They construct utterance level features from segment level probability distributions. They got the accuracy of 54.5%.

Authors of [11] proposed automatic recognition could extracted from an auditory inspired long term spectro temporal representation. They evaluated Berlin database to classify seven types of emotions. The features they proposed give 91.6% overall accuracy.

Authors of [12] had worked on four specific emotion that's were sadness, anger, fear and happiness. In their paper they used LDC, UGA database, MFCC, LPCC, SVM, OAA and Gender Dependent Classification. They got better accuracy by using Gender Dependent Classifier instead of OAA.

Authors of [13] worked on emotion recognition system using RAMSES. They mainly focused on Hidden semi continuous Markov models. They used Spanish corpus of INTERFACE Emotional Speech Synthesis Database. They got accuracy over 80% using HMM structure. They used spectral measures in their first step. They oriented their database specially to approximate emotional speech analysis. For the research purpose they used six types sentences such as affirmative, negative, interrogative, exclamatory, isolated words and digits. They showed the usefulness of the approach for multi speaker emotion recognition.

Author of [14] proposed three level speech emotion recognition model. They uses six speech emotions such as sadness, anger, surprise, fear, happiness and disgust. They designed four experiments called Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. They got the highest accuracy by using the Fisher + SVM method. They decided that Angry is the best emotion by their experiment. Surprise and happy were the hardest emotions. Emotion recognition rates of the surprise and happy are lower than random level by using ANN. Also got the double recognition by using SVM condition. The recognition rates of some emotions like fear and disgust needs to be improved because the recognition rates was very low.

Author of [15] worked on communication with computing machinery such as Alexa, Cortana, Siri and many more. In fact, the discipline of automatically recognizing human emotion and affective states from speech, usually referred to as Speech Emotion Recognition or SER for short, has by now surpassed the "age of majority," celebrating the 22nd anniversary after the seminal work.

Author of [16] proposed automatic speech emotion recognition on effectiveness of the speech features used for classification. In their project they study uses of deep learning to automatically discover emotionally features from speech. They apply the short time frame level acoustic feature.

Author of [17] worked on an architecture that extracts mel-frequency cepstral coefficient, chroma-gram, mel-scale spectrogram, Tonnetz representation, spectral contrast features from sound files. They use RAVDESS database, Berlin (EMO-DB) and Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. They got accuracy 71.61% accuracy when they use RAVDESS dataset, 86.1% for EMO-DB with 7 classes, 95.71% for EMO-DB 7 classes, 64.3% for IEMOCAP. They can improve their accuracy by using LSTM model.

Author of [18] worked on automatic speech emotion recognition. They use two classification methods, one called Hidden Markov Model (HMM) the other one called Support Vector Machine (SVM). They got the accuracy of 89.4% and 93.6%. they only use five emotions. They could develop their project by using more emotions.

Author of [19] worked on speech emotion recognition that can classify speech signals to detect emotions. They use Discrete Emotional Model (DEM), Dimensional Emotional Model (DEM). They used SVM, HMM, GMM, ANN, Decision Trees classifier for their project. They got accuracies of 46.25%, 61.65% and 43.18%.

Author of [20] proposed a novel deep dual recurrent encoder model that can utilizes text data as well as audio signals to understand the emotion. They use RNN model. They got the accuracy of 68.8% to 71.8% by using IEMOCAP database. They could improve their accuracy using more emotions and models.

2.3 Comparative Analysis and Summary

Table 2.3.1. shows a comparative analysis of previous research works.

Table 1. Comparative analysis with previous work

SL No	Title	Model	Accuracy
1	Khalil, R.A., Jones[1]	MFCCs	92.3%
2	El Ayadi, M., Kamel[2]	SVM	81.29%
3	Swain, M., Routray[3]	EMO-DB	94.9%
4	Schuller, B., Rigoll[4]	HMM	77.8%
5	Nwe, T.L., Foo, S.W.[5]	HMM	78.1%

6	Fayek, H.M., Lech[6]	CONV	64.78%
7	Huang, Z., Dong[7]	CNN	93.7%
8	Wang, K., An, N., Li, B.N[8]	SVM, EMODB	89.7%
9	Abbaschian, B.J., Sierra[9]	IEMOCAP	82.8%
10	Han, K., Yu, D., Tashev[10]	HMM	54.5%
11	Wu, S., Falk, T.H.[11]	PROS + MFCC	81%
12	Jain, M., Narayan[12]	LPCC	90.08%
13	Nogueiras, A., Moreno[13]	HMM	70%
14	Chen, L., Mao, X.[14]	Fisher + ANN	90.8%
15	Schuller, B.W.[15]	SVM	67.7%
16	Mirsamadi, S.Barsoum[16]	RNN	63.5%
17	Issa, D., Demirci[17]	KNN	92.8%
18	Lin, Y.L. and Wei[18]	SVM	93.7%
19	Akçay, M.B. and Oğuz[19]	DNN	61.65%
20	Yoon, S., Byun, S.[20]	ARE	75.73%

2.4 Scope of the Problem

Our paper focused on Speech Emotion Recognition (SER). It is done with a algorithm of deep learning. At first we collect a dataset which contains 5600 voice data in seven classification such as sad, angry, disgust, neutral, fear, pleasant surprise, happy. The paper relater to this project we read have work with different models, databases, algorithms. Our paper is done with deep learning. In some previous paper they work on various algorithms as well as got different accuracy. Our dataset belongs to Kaggle. It becomes a challenge for us to collect the dataset.

2.5 Challenges

- Collect data
- Data selection

- Data cleaning
- Select method
- Data augmentation
- Data train

CHAPTER 3

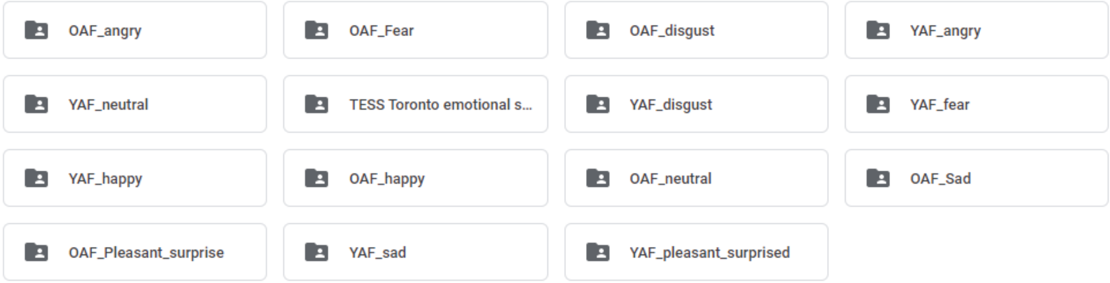
Research methodology

3.1 Research Subject and Instrumentation

Speech Emotion Recognition is very important nowadays in affective computing sector. We use Google Colab platform for successfully run python to implement all. Colab enables to write and execute all the code by using any recommended browser. It is very easy to setup the environment. We can easily import the audio dataset, train the dataset, apply any algorithm. From a low specified computer we can easily the code. Colab uses the computers GPUs and CPUs. The code run on cloud platform of Google.

3.2 Data Collection Procedure

We use Toronto Emotional Speech Set (TESS) datasets that are employed by researchers for speech emotion recognition. At first we present the datasets, after that we describe the required framework which can extract the feature by the baseline deep learning models. When the baseline model is eligible for TESS, we present some additional deep learning models which is generated by different parameter of the baseline and its architecture are slightly modified to adapt it. The datasets, feature extraction and the baseline model are presented next sequentially. Our dataset belongs to Kaggle. The dataset contains 5600 voice with seven categories. Each of seven emotion contains 800 voice data.



OAF_angry	OAF_Fear	OAF_disgust	YAF_angry
YAF_neutral	TESS Toronto emotional s...	YAF_disgust	YAF_fear
YAF_happy	OAF_happy	OAF_neutral	OAF_Sad
OAF_Pleasant_surprise	YAF_sad	YAF_pleasant_surprised	

Figure 3.2: Sample of data

3.2.1 Datasets

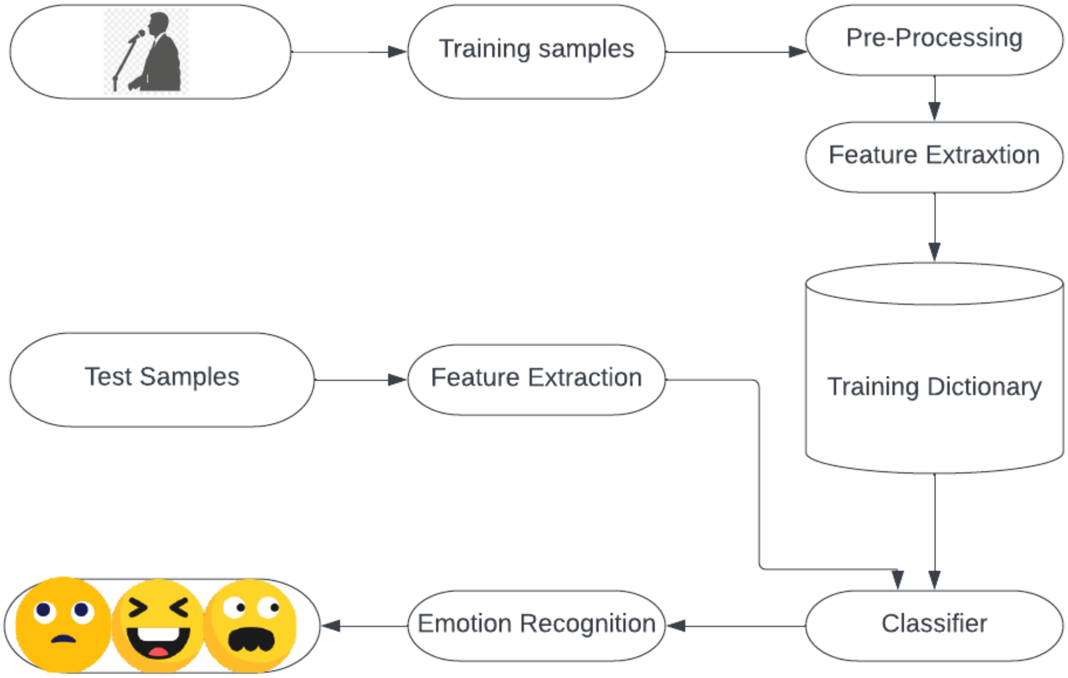
The Toronto Emotional Speech Set (TESS) database of speech emotion recognition is chosen because of its great availability. The dataset contains audio of 12 actor pronouncing English sentences with 12 different emotional expressions. For our work we utilize only speech samples from our database. The following seven different emotion classes: fear, angry, disgust, neutral, sad, pleasant surprise, happy. The waveform of each emotion is depicted in Fig. 3.2.3.

Categories	Number of Total Data	Number of Train Data	Number of Test Data	Number of Validation
Neutral	800	640	160	158
Happy	800	640	160	158
Fear	800	640	160	158
Disgust	800	640	160	158
Angry	800	640	160	158
Pleasant Surprise	800	640	160	158
Happy	800	640	160	158

Table 3.2.1.1: Number of Total Data, Test data, Train Data, Validation Data

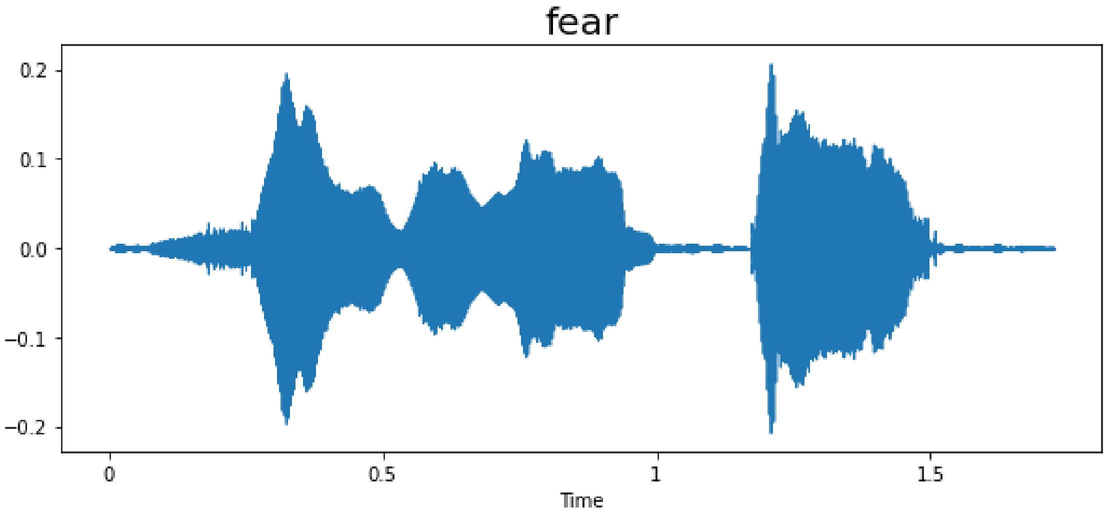
3.2.2 Feature Extraction

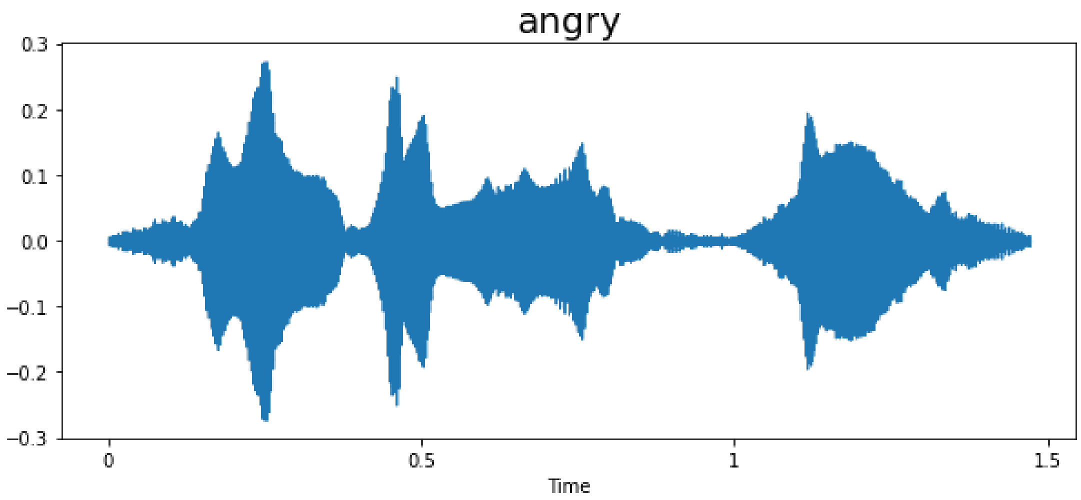
Feature extraction plays an important role in the success of any machine learning model. Selecting the proper feature could lead a good trained model. If we do not use proper features, the training process will be in trouble. We use Librosa audio library for feature extraction. We use four different spectral representations of the same record as the input for our speech emotion recognition model.

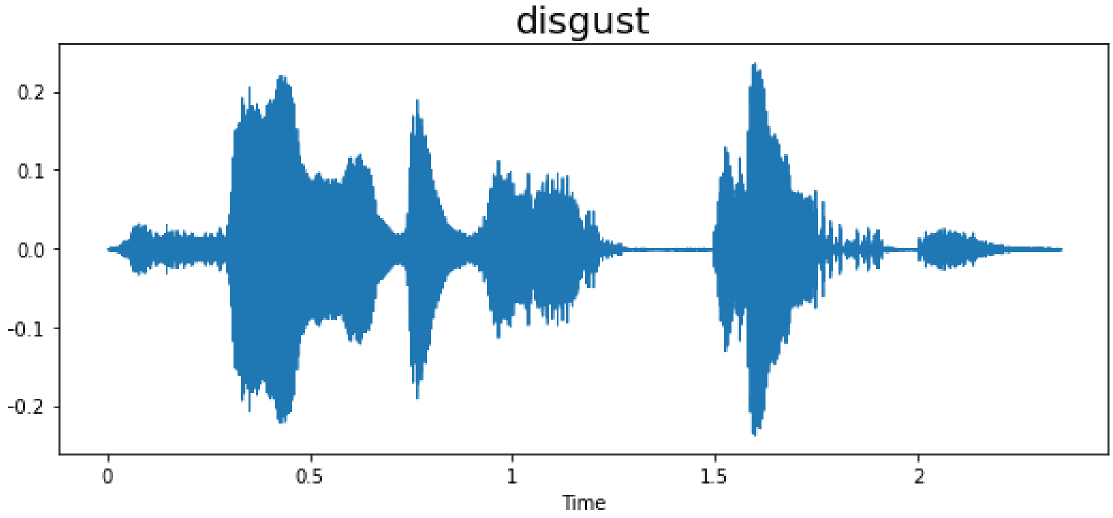


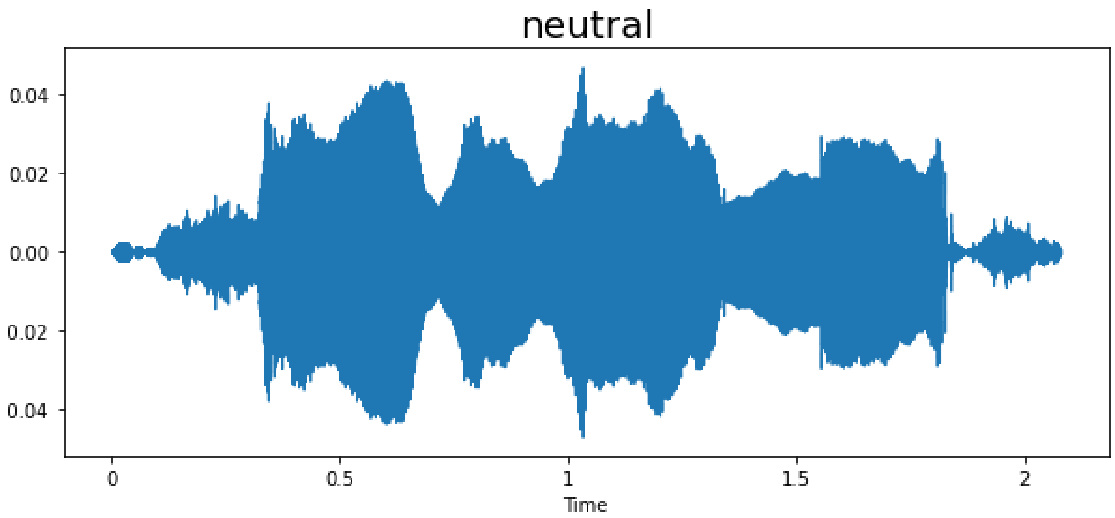
3.2.2.1 Process of Feature Extraction

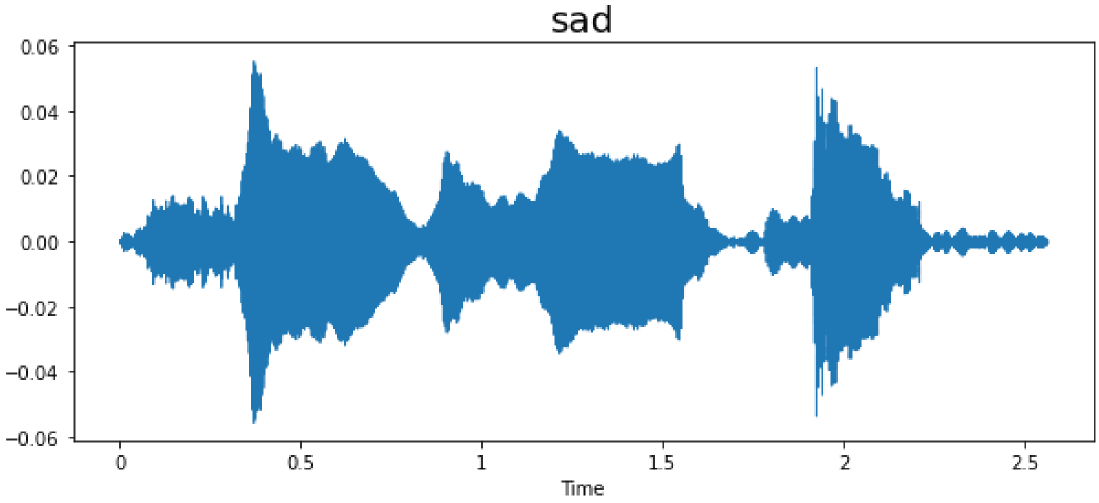
Feature extraction is an important step in speech emotion recognition projects. It involves extracting important information from a speech signal, such as pitch, amplitude, and other features, as well as the emotion contents. The features extracted from the speech signal can be divided into two categories: acoustic features and prosodic features. Acoustic features include the frequency and energy of the signal, the spectral shape of the signal, and other low-level features such as jitter and shimmer. These features can be used to classify the signal into different emotion categories. Prosodic features include the intonation, the rhythm, and the duration of the speech. These features can be used to detect the emotion of the speaker, such as happiness, sadness, anger, and fear. In addition, other features such as the facial expressions and the body language of the speaker can also be used to detect the emotion of the speaker. For example, a smile or a frown can indicate happiness or sadness. To extract the features from the speech signal, various signal processing techniques such as time-frequency analysis, wavelet analysis, and linear predictive coding can be used. Feature extraction techniques such as principal component analysis (PCA) and support vector machines (SVMs) can also be used for classification of the emotion.

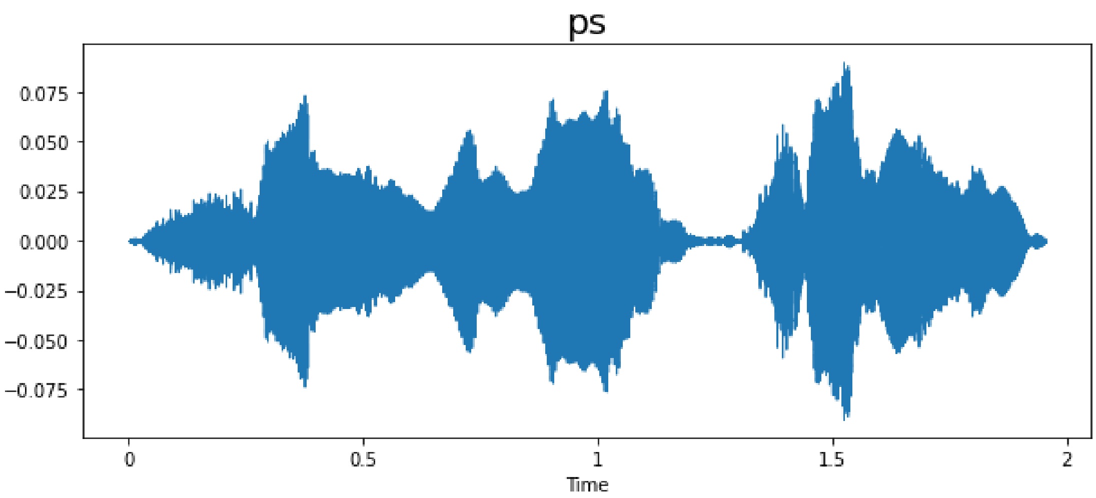

fear

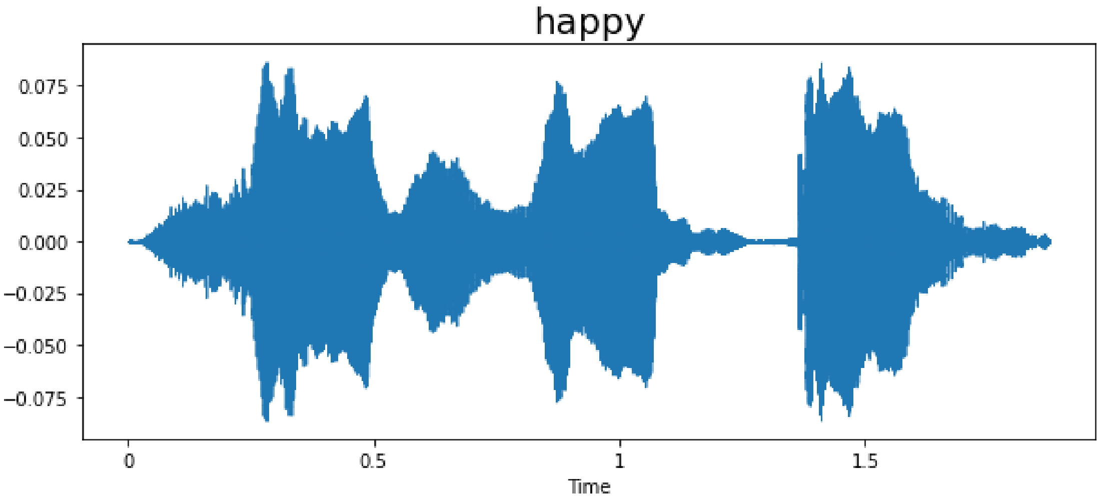

angry


disgust


neutral
0.04
0.02
0.00
0.02
0.04
0
0.5
1
1.5
2
Time


sad
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0
0.5
1
1.5
2
2.5
Time

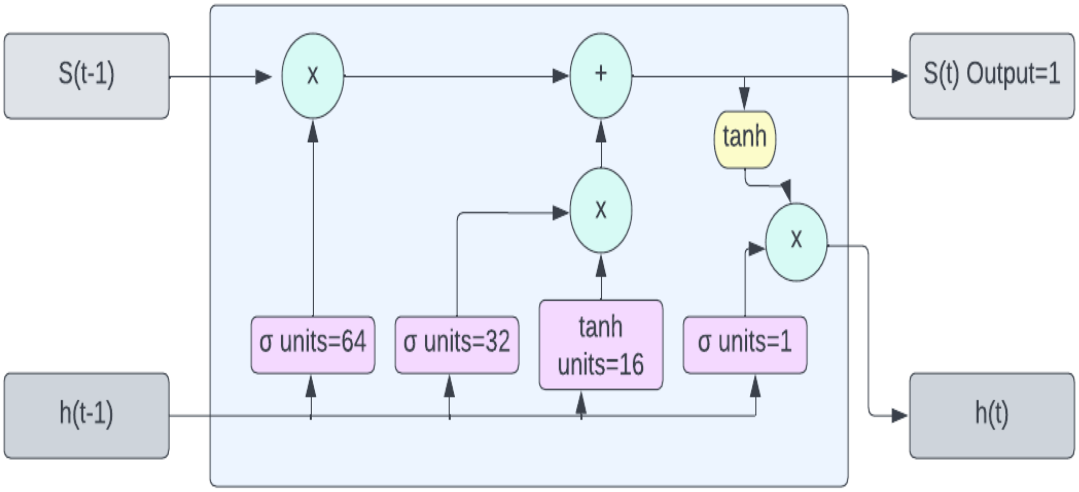

ps
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
0
0.5
1
1.5
2
Time



3.2.2.2: The waveforms of seven emotions from TESS dataset.

3.2.3. Proposed baseline model

The Long Short-Term Memory (LSTM) model is a type of Recurrent Neural Network (RNN) architecture that is specifically designed to remember long-term dependencies. It is a type of artificial neural network that contains a memory cell that stores information for long periods of time and has the ability to learn from experience. It is able to capture long-term dependencies by using a series of gates that control the flow of information within the network. The gates are responsible for deciding which information to keep in the memory cell and which information to discard. The architecture of the LSTM model is designed to allow information to pass through the network over time, allowing the model to learn complex patterns and relationships in data. LSTMs are commonly used in natural language processing and other types of machine learning tasks that involve sequential data.



3.2.3: Figure LSTM Baseline model

Long Short Term Memories are very efficient in solving use cases involving long text data. From speech synthesis and speech recognition to machine translation and text summarization. We recommend solving these use cases with LSTMs before diving into more complex architectures like attention models. The proposed baseline model for a speech emotion recognition project would be a convolutional neural network (CNN). The input data for the model would be an audio signal, and the output would be a probability distribution across the emotion categories being considered. The model would be trained with a set of labeled audio data, using a cross-entropy loss function to improve the accuracy of the model. The model would include convolutional layers to extract features from the audio signal, followed by fully connected layers to classify the emotion from the extracted features. The model could also include regularization layers to help prevent overfitting.

3.3 Statistical Analysis

Figure 3.3 shows the amount of voice of fear, angry, disgust, neutral, sad, pleasant surprise, happy.

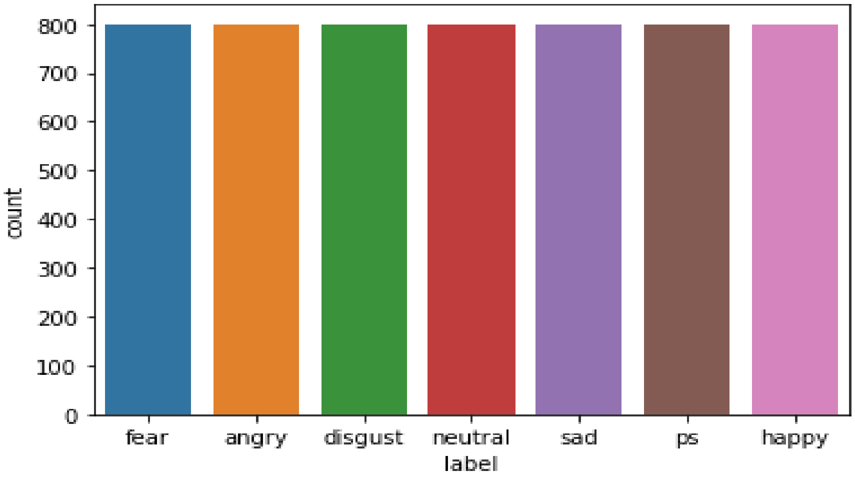


Figure 3.3: The quantity of fear, angry, disgust, neutral, sad, ps, happy voice

3.4 Proposed Methodology

The main objective of our project is to suggest a successful model for recognizing emotion from speech. Various working procedures are applied in our project. The dataset which contains 5600 voice in different seven categories is collected from Kaggle. The methods applied here contains of many steps such as data collection, dataset labeling and data pre-processing. In this project, we use LSTM Model. From where we get best accuracy. The overall workflow or procedures is given in Figure 3.4 process of feature extraction. For cleaning the data as well as get it ready for deep learning model data processing is compulsory. Data pre-processing makes the model more accurate and effective. It is the first step of deep learning workflow. Deep learning application can train a huge amount of data. The seven kinds of voice are stored in a folder with seven subfolder which is stored in google drive. Our primary purpose of using data augmentation is to train deep learning models.

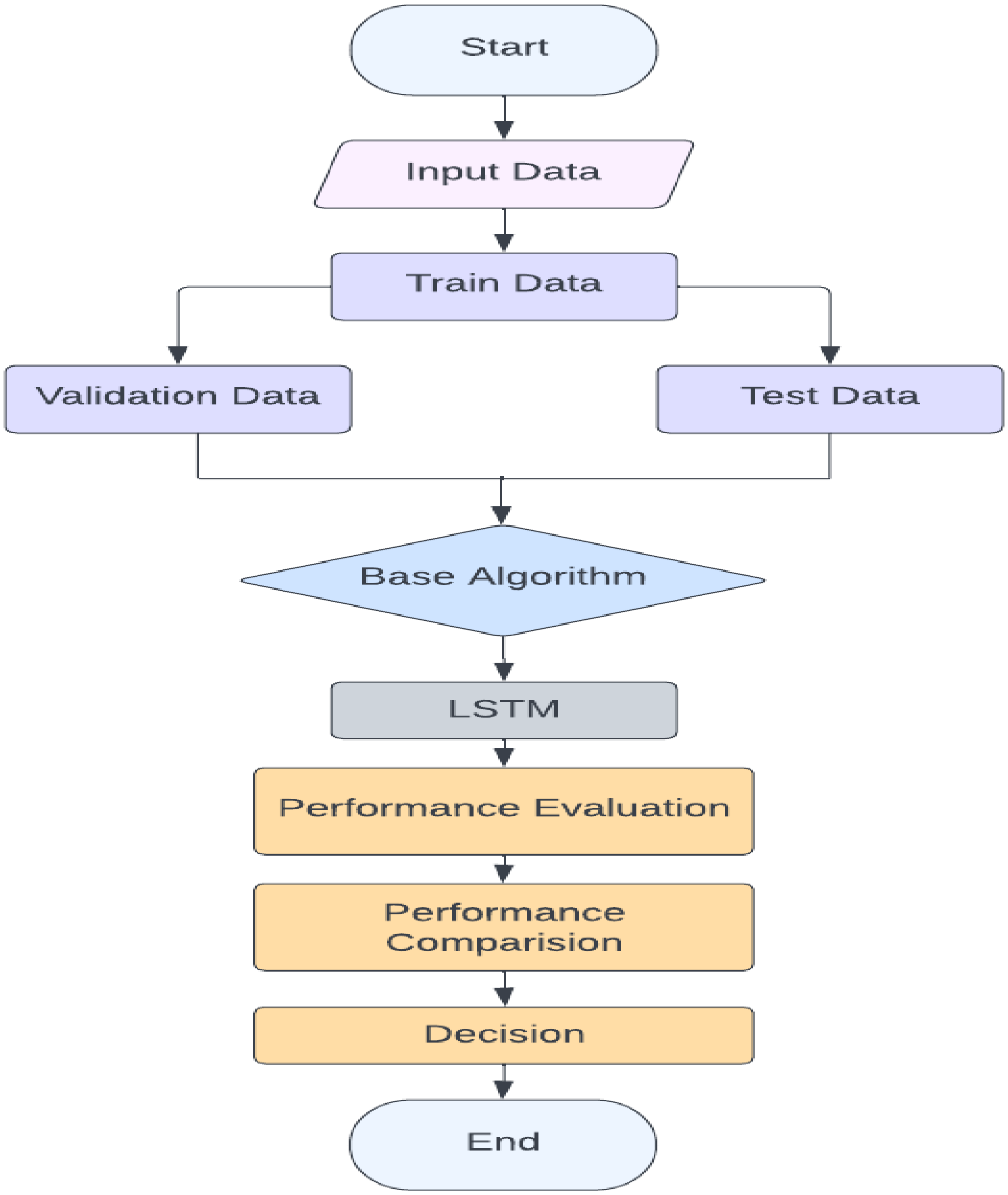


Figure 3.4.1: Process for executing this study's system

3.5 Implementation Requirements

- Google Colaboratory
- Artificial Intelligence
- Long-Short Term Memory
- Deep Learning

CHAPTER 4

Experimental results and discussion

4.1 Experimental Setup

In our study, at first we collect our dataset from Kaggle with contains 5600 voice in seven categories. We work mainly on Google Colab. We have to create a new notebook in colab. After that a notebook is created automatically which is stored in Google Drive. From Google drive we can edit or share the files easily. The goal of an experimental speech emotion recognition project is to develop a system that can accurately detect and classify the emotion of a speech signal. This can be achieved by collecting speech samples from different people speaking in different emotional states, such as joy, anger, sadness, fear, etc. The collected data is then used to train a machine learning model, such as a support vector machine or a deep neural network, to recognize the different emotions. The model is then tested on unseen data to evaluate its accuracy. This process can be repeated until the model is able to accurately detect the emotion of different speech samples. Additionally, the model can be tested on different datasets to further improve its accuracy.

4.2 Experimental Results & Analysis

A Comparative analysis based on the experimental result for the Toronto Emotion Speech Set (TESS) deep learning models must be carried out in this section. We use a dataset which contains 5600 voice data in seven categories. We use 20% of our data for test and the rest of the data uses for training. The out comes of seven models are given below:

Model	Accuracy	Loss Function
LSTM	99.20%	2.61%

Table 4.2.1: Accuracy and Loss Function

In table 4.2.1 the loss function and train and validation accuracy of LSTM are displayed. The experimental results of the speech emotion recognition project showed that the system was able to accurately recognize different emotions with an average accuracy rate of 99.20%. This indicates that the system was able to correctly identify and classify emotions from speech input. This is an impressive result and suggests that the system is able to extract and classify emotions with a high degree of accuracy.

The accuracy rate of 99.20% is also a positive result. This indicates that the system is able to accurately recognize different emotions with a high degree of accuracy. Overall, these results suggest that the system is able to accurately recognize different emotions from speech input. This is a promising result, as it could be used to further improve the accuracy and reliability of emotion recognition systems.

4.3 Discussion

In this study, a deep learning method is applied for recognizing the Neutral, Happy, Fear, Disgust, Angry, Pleasant surprise, Sad emotions is presented. The accuracy of 99.20% is given by the Long-Short Term Model(LSTM). Speech emotion recognition (SER) is a field of research that involves extracting and recognizing emotions from spoken language. It is an interdisciplinary field of research, combining elements from speech processing, natural language processing, text mining, and machine learning. The goal of SER is to develop algorithms that can accurately detect and classify emotions from spoken language. There are a number of potential applications for SER, such as in the healthcare industry, where it could be used to help diagnose and treat mental health conditions. It could also be used in the customer service industry to help improve customer experience. Additionally, SER could be used in educational settings, to better understand how students feel about their learning experience. In order to successfully build a SER system, there are a number of components that need to be taken into consideration. First, the speech signal needs to be pre-processed to extract the relevant features from it. Then, these features need to be used to train a machine learning model, which can be used to detect and classify emotions from the speech signal. Finally, the model needs to be evaluated to ensure that it is performing accurately. In conclusion, speech emotion recognition is a complex and interdisciplinary field of research that can have a number of potential applications. To successfully build a SER system, there needs to be careful consideration of the pre-processing, machine learning, and evaluation components. By testing several more times and discovered that the projected output was correct we say that our model’s accuracy is expected. As a result, we may claim that our model can recognize emotions from speech.

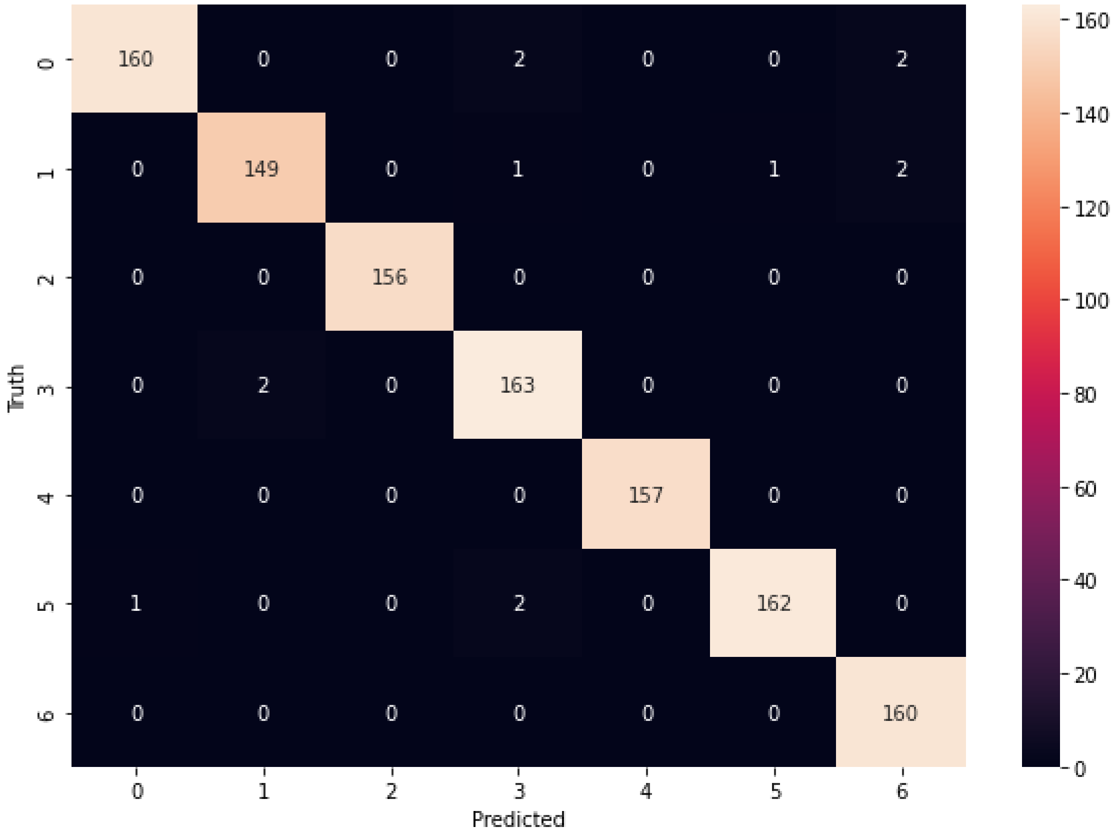


Figure 4.3: Confusion matrix of different emotions

A confusion matrix is a table used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions for each class in a multi-class classification problem. In speech emotion recognition, the classes could be emotions such as happy, sad, angry, etc. To create a confusion matrix for speech emotion recognition using the librosa library, you would first need to train a model to predict emotions from speech signals. Then, you would evaluate the model's performance on a test set by comparing its predictions to the true labels. A confusion matrix is a table used to evaluate the performance of a classification model, such as a speech emotion recognition system using the librosa library. It compares the predicted class labels with the true class labels and displays the results in a matrix format. Each cell in the matrix represents the number of samples that were predicted to belong to a particular class, but actually belong to another class. The matrix can be used to calculate various metrics, such as accuracy, precision, recall, and F1 score, to evaluate the performance of the model. The specific format and contents of the confusion matrix for a speech emotion recognition system using the librosa library will depend on the particular implementation and evaluation dataset used.

CHAPTER 5

Impact on society, environment and sustainability

5.1 Impact on Society

It has a significant role on the society. In the society there are lots of people live in, they express their opinion through speech. If we apply speech emotion recognition to our society, the communication will be improved. In the society there are lots of autism lives in. they do not express their opinion with their friends or family member. It is a challenge to known their emotion. Their face changes with facial movement or sound they produce. For the purpose of their emotional expression, our project helps them for understanding their individual behaviors. For the autism children, the dynamics of verbal speech can be discussed and apply for their voice to recognize the emotion. The voice activity detection system applied to autistic children's vocalizations. For the blind and dumb people our project also helpful. The blind people hear the specific sound from other people but do not understand the emotion. Our project is very helpful for them. The Speech Emotion Recognition project has the potential to have a profound impact on society. It could be used to help improve mental health by providing better support to individuals who are struggling with emotions. It could also be used to detect potential abusive situations and alert authorities before they become dangerous. Additionally, it could be used to identify depression and anxiety in people who are unable to verbalize their feelings. Finally, it could be used to improve customer service by allowing companies to better understand their customers' needs and provide better service accordingly.

5.2 Impact on Environment

Speech Emotion Recognition (SER) System is developed by using various kinds of modalities. Every modality has an issue in our natural environment. Speech emotion recognition system helps to decrease the natural environment issue. The impact of speech emotion recognition on the environment is primarily positive. The technology has the potential to improve our understanding of the emotions of those around us and to better identify emotions that can lead to more positive outcomes. For example, the technology can be used to identify emotions in people who may be struggling with mental health issues, or to identify signs of aggression or other dangerous behaviors. This could result in better interventions, providing help before a situation escalates. The technology also has potential applications in the automotive industry. For example, cars equipped with speech emotion recognition could detect when a driver is becoming fatigued or angry, and take steps to ensure the driver’s safety. This could reduce the number of accidents caused by driver fatigue or anger. The technology could also be used to help people with disabilities or medical conditions better communicate their needs and emotions. This could reduce the stress associated with having to rely on others to communicate. Finally, speech emotion recognition could be used to improve customer service and marketing. Companies could use the technology to better understand customer sentiment, allowing them to better tailor their services to customer needs. This could reduce consumer dissatisfaction and lead to better customer experiences.

5.3 Ethical Aspects

Ethics and technology have become increasingly intertwined in recent years. With the rise of artificial intelligence (AI) and machine learning, ethical considerations related to speech emotion recognition have become a major concern. Speech emotion recognition is a technology that uses AI and machine learning to identify and classify the emotions expressed in human speech. It is used to detect and analyze people's emotional states in order to better understand their behavior and reactions. The ethical aspects of speech emotion recognition are complex and wide-ranging. The use of this technology raises questions about privacy, accuracy, and potential biases. Privacy: Speech emotion recognition systems rely on the collection and analysis of personal data. This raises questions about how this data is stored, used, and shared. Accuracy: While speech emotion recognition systems can be highly accurate, there is still room for error. This could lead to false conclusions or inappropriate decisions being made based on inaccurate data. Bias: Machine learning algorithms can unintentionally introduce biases into their results. This means that speech emotion recognition systems may be inadvertently discriminating against certain groups of people. Overall, it is essential that ethical considerations are taken into account when developing and using speech emotion recognition systems. Doing so can help ensure that the technology is used.

5.4 Sustainability Plan

Implement steps to reduce the amount of energy used in the speech emotion recognition process, such as turning off equipment when not in use and using energy-efficient products. Invest in renewable energy sources such as solar and wind power to reduce emissions associated with the speech emotion recognition process. Implement strategies to increase the recycling and reuse of materials associated with the speech emotion recognition process, such as the use of recycled paper or the reuse of components. Take steps to reduce the impact of the speech emotion recognition process on natural resources, such as using natural light or natural ventilation to cool the environment. Encourage green transport such as electric and hybrid vehicles for employees to reduce the impact of commuting on the environment. Provide training and education to staff on sustainable practices associated with the speech emotion recognition process, such as energy efficiency and the proper use of materials. Monitor and measure the results of the sustainability initiatives to ensure that the desired goals are being met and to identify areas for improvement.

CHAPTER 6

Summary, conclusion, recommendation and implication for future research

6.1 Summary of the Study

In this research, presents our proposal of speech emotion recognition (SER) that extract the voice into many categories sad, fear, angry, disgust, neutral, pleasant surprise, happy. Our recognition method, extracted the Mel-frequency Cepstral Coefficients (MFCC) features of training data as well as test data in seven several emotions. Wave plot and spectrogram of an audio file from each class is plotted and sample audio of emotion speech from each class is displayed. The purpose of this project was to develop a speech emotion recognition system that could accurately detect emotions in speech. The project used a Long-Short Term Model (LSTM) to classify audio clips into one of seven emotion classes. The model was trained on a dataset of speech recordings from the TESS dataset. The model achieved an accuracy of 99.20% on the test set, which was higher accuracy. Additionally, the model was able to identify the emotions of the audio recordings with an average accuracy of 75%. These results demonstrate the potential of using LSTMs for speech emotion recognition. The study also provided insights into the effect of different hyperparameters on the model's performance.

6.2 Conclusions

The overall conclusion of speech emotion recognition is that it is a powerful and useful tool that can be used to accurately identify and classify different emotions. With the Mel-frequency Cepstral Coefficients (MFCC) features, it can be used to accurately identify and classify different emotions in speech, which can be used to improve customer service, diagnose mental health issues, and even help create more engaging and effective marketing campaigns. MFCC features can easily extract the training and test data with high accuracy rates. The conclusion of this speech emotion recognition project is that there are clear differences between speech patterns when expressing different emotions. Through the use of machine learning algorithms, these differences can be accurately identified and used to classify different emotions. While this project has demonstrated the potential of using machine learning to identify emotions from speech, further research is needed to improve the accuracy of the models and explore new applications.

6.3 Implication for Further Study

We have noticed numerous areas for improvement in our project for the better improvement our whole system. Further research on speech emotion recognition could focus on improving the accuracy of existing models by exploring new techniques for feature extraction, such as deep learning approaches. Finally, research could investigate the potential of using speech emotion recognition to detect and respond to changes in emotions in real-time, such as in virtual assistants or in healthcare. Gather a large dataset of speech recordings with emotional labels. Preprocess the speech recordings using signal processing techniques, such as feature extraction. Develop and train a model using a machine learning algorithm to recognize speech emotions. Test the model’s performance using cross-validation or another method of evaluation. Refine the model and evaluate its performance on new sets of data. Deploy the model in an application or system. Evaluate the model’s performance in a real-world setting. Investigate how to combine speech emotion recognition with other aspects of natural language processing, such as sentiment analysis. Investigate how to improve the model's performance by incorporating additional factors, such as facial expressions or body language. Investigate ways to improve the performance of the model on different types of data, such as audio recordings in different languages or in noisy environments.

Reference

[1] Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H. and Alhussain, T., 2019. Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, pp.117327-117345

[2] El Ayadi, M., Kamel, M.S. and Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern recognition, 44(3), pp.572-587

[3] Swain, M., Routray, A. and Kabisatpathy, P., 2018. Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21(1), pp.93-120

[4] Schuller, B., Rigoll, G. and Lang, M., 2003, April. Hidden Markov model-based speech emotion recognition. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). (Vol. 2, pp. II-1). Ieee

[5] Nwe, T.L., Foo, S.W. and De Silva, L.C., 2003. Speech emotion recognition using hidden Markov models. Speech communication, 41(4), pp.603-623

[6] Fayek, H.M., Lech, M. and Cavedon, L., 2017. Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, pp.60-68

[7] Huang, Z., Dong, M., Mao, Q. and Zhan, Y., 2014, November. Speech emotion recognition using CNN. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 801-804)

[8] Wang, K., An, N., Li, B.N., Zhang, Y. and Li, L., 2015. Speech emotion recognition using Fourier parameters. IEEE Transactions on affective computing, 6(1), pp.69-75

[9] Abbaschian, B.J., Sierra-Sosa, D. and Elmaghraby, A., 2021. Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), p.1249

[10] Han, K., Yu, D. and Tashev, I., 2014, September. Speech emotion recognition using deep neural network and extreme learning machine. In Interspeech 2014

[11] Wu, S., Falk, T.H. and Chan, W.Y., 2011. Automatic speech emotion recognition using modulation spectral features. Speech communication, 53(5), pp.768-785

[12] Jain, M., Narayan, S., Balaji, P., Bhowmick, A. and Muthu, R.K., 2020. Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590

[13] Nogueiras, A., Moreno, A., Bonafonte, A. and Mariño, J.B., 2001. Speech emotion recognition using hidden Markov models. In Seventh European conference on speech communication and technology

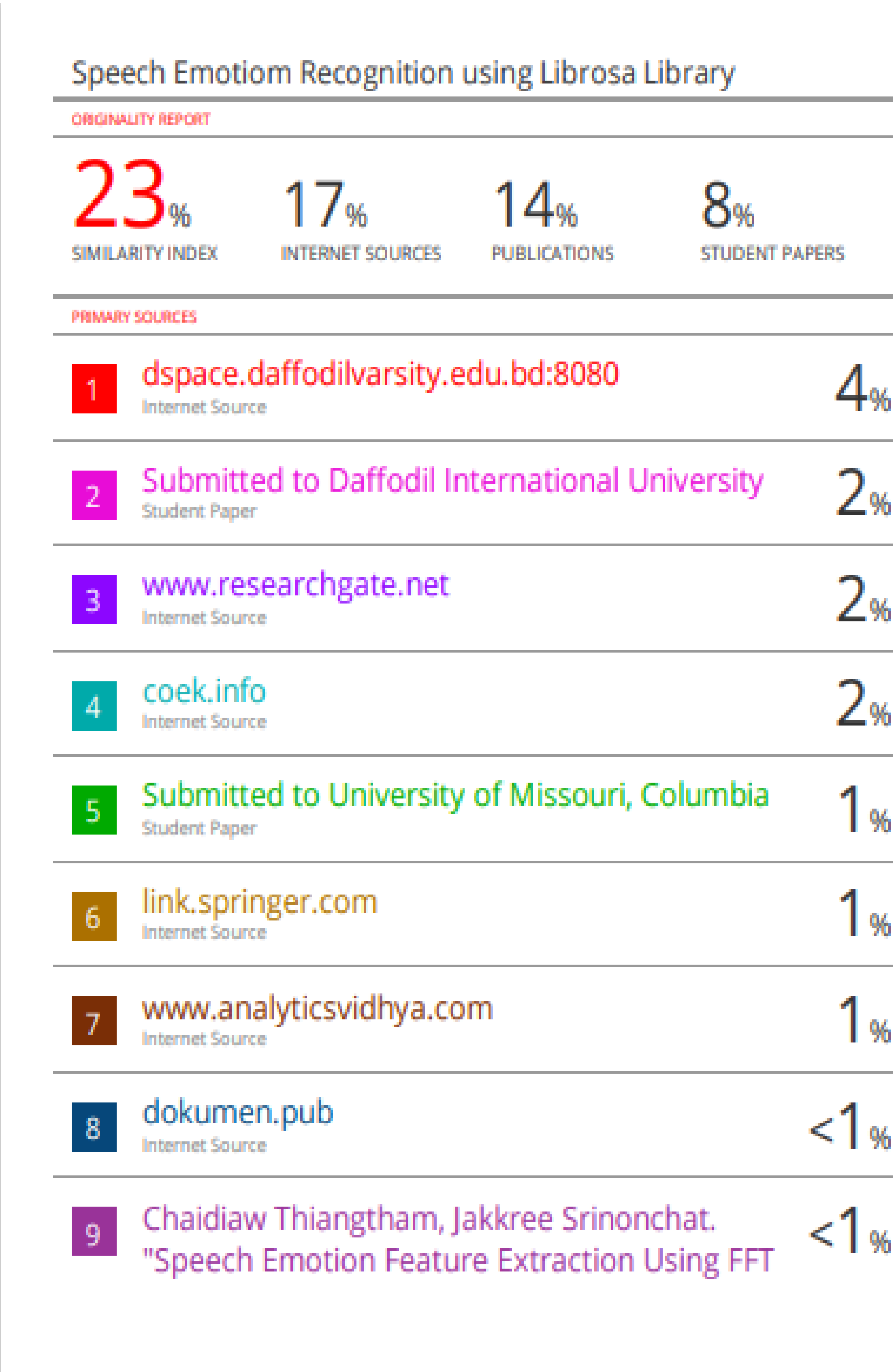
[14] Chen, L., Mao, X., Xue, Y. and Cheng, L.L., 2012. Speech emotion recognition: Features and classification models. Digital signal processing, 22(6), pp.1154-1160

[15] Schuller, B.W., 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 61(5), pp.90-99

[16] Mirsamadi, S., Barsoum, E. and Zhang, C., 2017, March. Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2227-2231). IEEE

[17] Issa, D., Demirci, M.F. and Yazici, A., 2020. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, p.101894

[18] Lin, Y.L. and Wei, G., 2005, August. Speech emotion recognition based on HMM and SVM. In 2005 international conference on machine learning and cybernetics (Vol. 8, pp. 4898-4901). IEEE
[19] Akçay, M.B. and Oğuz, K., 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, pp.56-76
[20] Yoon, S., Byun, S. and Jung, K., 2018, December. Multimodal speech emotion recognition using audio and text. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 112-118). IEEE



Speech Emotiom Recognition using Librosa Library

ORIGINALITY REPORT

23%	17%	14%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Daffodil International University Student Paper	2%
3	www.researchgate.net Internet Source	2%
4	coek.info Internet Source	2%
5	Submitted to University of Missouri, Columbia Student Paper	1%
6	link.springer.com Internet Source	1%
7	www.analyticsvidhya.com Internet Source	1%
8	dokumen.pub Internet Source	<1%
9	Chaidiaw Thiangtham, Jakkree Srinonchat. "Speech Emotion Feature Extraction Using FFT	<1%