

SPEECH ANALYSIS SYSTEM USING SPEECH RECOGNITION AND EMOTION RECOGNITION METHODS

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology degree in Information Technology

By

EVANGELIN E (Reg.No - 39120030)



DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600119**

MAY- 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

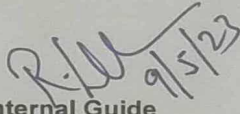
Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

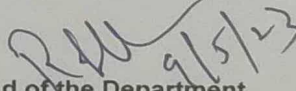
DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Evangelin E** (Reg.No - 39120030) who carried out the Project Phase-2 entitled "SPEECH ANALYSIS SYSTEM USING SPEECH RECOGNITION AND EMOTION RECOGNITION METHODS" under my supervision from January 2023 to May 2023 .

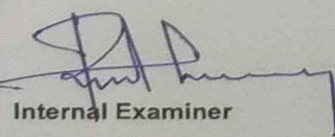

Internal Guide

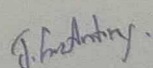
Dr. R. SUBHASHINI, M.E., Ph.D.


Head of the Department

Dr. R. SUBHASHINI, M.E., Ph.D.

Submitted for Viva voce Examination held on 10/05/2023

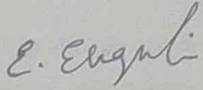

Internal Examiner


External Examiner

DECLARATION

I, Evangelin E(Reg.No - 39120030), hereby declare that the Project Phase-2 Report entitled "SPEECH ANALYSIS SYSTEM USING SPEECH RECOGNITION AND EMOTION RECOGNITION METHODS" done by me under the guidance of Dr. R. SUBHASHINI M.E., Ph.D. is submitted in partial fulfillment of the requirements for the award Bachelor of Technology degree in INFORMATION TECHNOLOGY.

DATE: 10/05/2023


SIGNATURE OF THE CANDIDATE

PLACE: Chennai

EVANGELIN E

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr.R.SUBHASHINI M.E., Ph.D.**, Head of the Department, Department of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.R.SUBHASHINI M.E., Ph.D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

ABSTRACT

As automation's are increasing in this developing world, the need for a machine to understand human speech and their emotions has become an important feature , and also in the field of customer services providers like call centers, which contains large collection of voice recordings collected in various ways as feed-backs or complains has a need to be analyzed to provide useful information. Many of those service providers are started using speech analysis methods to fetch those useful in-formations from the speech recordings such as the demographic characteristic and the emotional state of the speaker and the verbal content from the speech. Speech analysis, an important topic of interest in the fields of psychology , neuroscience and artificial intelligence is the process of extracting meaningful in-formations from the audio signals such as identity of the speaker , emotion state of the speaker and content from the speech using speech recognition software and methods. This paper proposes a speech analysis system to analyse the input speech using the speech recognition methods and to predict the gender, age of the speaker and emotional state of the speaker. Four models are trained : (i) Gender Prediction model based on the acoustic feature to predict gender as Male, Female or Other, (ii) Age Prediction model based on the acoustic feature to predict age group as Teens, Twenties, Thirties, Forties, Fifties, Sixties and Seventies, (iii) Speech Emotion Recognition model based on acoustic feature to predict Eight class of emotions such as Anger, Sad, Fear, Happy, Surprise, Neutral, Calm and Disgust and (iv) Text Emotion Recognition model based on linguistic feature to predict Seven class of emotions such as Anger, Sad, Fear, Joy, Surprise, Neutral and Disgust.

TABLE OF CONTENTS

Chapter No.	TITLE	Page No.
	ABSTRACT	iv
	LIST OF FIGURES	vii
1	INTRODUCTION	1
2	LITERATURE SURVEY	3
	2.1 Inferences	3
	2.1.1 Literature Survey	3
	2.2 Open problems in Existing System	5
3	REQUIREMENTS ANALYSIS	7
	3.1 Feasibility Studies/Risk Analysis of the Project	7
	3.2 Software Requirements Specification Document	8
	3.2 System Use Case	12
4	DESCRIPTION OF PROPOSED SYSTEM	14
	4.1 Selected Methodology or process model	14
	4.2 Architecture / Overall Design of Proposed System	15
	4.3 Description of Software for Implementation and Testing plan of the Proposed Model/System	18
	4.4 Project Management Plan	20
5	IMPLEMENTATION	21
	5.1 Development and Deployment Setup	21
	5.2 Modules Implemented	22

	5.3	Algorithm	28
6		RESULTS AND DISCUSSION	30
7		CONCLUSION	32
	7.1	Conclusion	32
	7.2	Future Scope	33
		REFERENCES	34
		APPENDIX	36
		A. SOURCE CODE	36
		B. SCREENSHOTS	49
		C. RESEARCH PAPER	54

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
3.1	Use Case Diagram of Speech Analysis System	13
4.1	Flow Diagram of Gender and Age Model	15
4.2	Flow Diagram of Speech Emotion Model	16
4.3	Flow Diagram of Text Emotion Model	16
4.4	Flow Diagram Diagram of Speech Analysis System	17
5.1	Statistical Representation of Gender and Age Dataset	23
5.2	Statistical Representation of Speech Emotion Dataset	25
5.3	Statistical Representation of Text Emotion Dataset	27
B.1	Data-set used for gender and age model	49
B.2	Speech Samples used in Gender and Age model	49
B.3	Features extracted from the speech sample for Gender	50
B.4	Classification report of Gender Model	50
B.5	Classification report of Age Model	50
B.6	Speech sampled dataset of Speech Emotion Model	51
B.7	Extracted features for Speech Emotion Model	51
B.8	Classification report of Speech Emotion Model	51
B.9	Prediction results of Speech Emotion Model	52
B.10	Labeled Dataset of Text Emotion Model	52
B.11	Classification Report of Text Emotion Model	52
B.12	Prediction Result of New Input	53

CHAPTER 1

INTRODUCTION

In this developing world, automated machines and applications are developed using various technologies like Artificial Intelligence and Machine Learning , Internet of Things etc.. those application mainly uses speech as their main mode of communication .Applications and Machines developed with voice activated assistance are becoming an popular usage among humans because the ability of a machines to understand, respond and communicate for the users commands makes them easy to access .The technologies and methods behind this voice activated assistance are Speech Recognition, Which is actually referred as Automatic Speech Recognition(ASR) is an type of artificial intelligence method that uses various algorithms and machine learning methods to analyse the human speech and convert it to text. Speech Recognition methods are not only used in developing an voice activated assistance but also in many fields like education to transcribe the audio and video recordings of the lectures and interviews to text for future use, in healthcare to transcribe medical notes, reports and other documents by reducing the need for manual transcription, also used in developing devices for disable people to help them with their disabilities by providing them easier way of accessibility to the devices and it is also mainly used in the customer services to automate call routing and customer identification as well as to analyse customer feedback and to improve service quality. When we talk about customer services , an enterprise providing services has to improve their services in order to maintain their profits, to improve their services an enterprise requires the response and feedback's about their services, those feedback's are collected from the customers in many ways, by conducting surveys, through social media platforms, using feedback forms for requesting their ratings and their opinions about the the services, through online review platforms and through call centers.Call Centers is a centralized office used to make and receive large volume of telephone calls , Call centers are commonly used by companies for customer service, telemarketing, technical support, and other business purposes.These call centers uses speech analytics as tool to analyse the voice recordings and to identify the reason for the call , the product mentioned and the mood of the caller.

Speech analysis is used to analyse the voice recordings and fetch useful information such as the identity of the speaker, emotional state of the speaker and content of the speech.In recent years, there has been a growing interest in using speech analysis for the prediction of demographic characteristics, such as gender and age, as well as for the prediction of emotional states.The prediction of demographic characteristics using speech analysis has become an active research area in recent years. Gender and age are two fundamental

demographic characteristics that can provide useful information for many applications, such as marketing research, healthcare, and law enforcement. The prediction of gender and age from speech signals can be achieved by analyzing the audio signal and extracting feature from it and these features can be used to train a machine learning model to predict gender and age. Emotion detection from speech has also become an important research area in recent years. Emotions play a critical role in human communication, and the ability to detect emotions from speech signals can have many practical applications, such as in mental health diagnosis, customer service, and education. Emotion detection from speech can be achieved by analyzing various acoustic features, such as pitch, intensity, and voice quality. In addition, lexical features, such as the use of certain words and language style, can also be used to predict emotions.

The aim of this project is to develop an speech analysis system which contain trained models to predict the age , gender and emotion of the speaker. The model to predict age and gender uses the acoustic characteristics of the speech to make the predictions where as two models are built to predict the seven class of emotions such as anger, sad, neutral, disgust, surprise, fear, and joy from the speech, one is to predict the emotion based on the acoustic feature and the other predicts based on the linguistic feature of the speech. The models that predicts gender, age and emotion based on the acoustic characteristics of the speech uses large data-set of speech samples, the acoustic features are extracted from these speech samples and are used as the data to train the model using machine learning and deep learning algorithms and the model to predict the emotion based on the linguistic characteristics of the speech is trained using a text documents that contains large number of sentences which expresses different emotions as the data-set and these sentences gets normalized by using various functions to remove the stop-words, to remove the punctuation's and trained using machine learning algorithms. These model uses supervised learning approach and large numbers of labeled data-sets to train the model. The performance of the model will be evaluated using various metrics, such as accuracy ,precision, recall and F1 score. And this speech analysis system is designed as an applications which will get the input from the user as an audio file that contains speech and displays the predicted results from the trained models as the output.

The proposed model has several potential applications. For example, it can be used in healthcare to detect the emotional state of patients with mental health disorders, such as depression and anxiety. It can also be used in customer service to improve the quality of interaction between customers and service providers. Moreover, it can be used in law enforcement to identify the demographic characteristics of suspects based on their speech signals.

CHAPTER-2

LITERATURE SURVEY

2.1. INFERENCES

As the automated voice activated technologies are increasing the need to analyse human speech without human interventions are also increasing. Many researches have been carried out in the field of speech analytic, such as models that are trained using popular Machine Learning and Deep learning algorithms to find the demographic characteristics like age and gender from the speech using their acoustic features and emotion state from the speech using both acoustic and linguistic features. Some of the insites from those researches are taken to improve our project.

2.1.1. LITERATURE SURVEY

The work of Saksham Goyal[1] , mentions that gender and age estimation from speech as an challenging research area for voice technology domains. And he discuss about the model built to predict the gender and age group from speech features extracted using various techniques such as Mel-Frequency Cepstral Coefficient, formants, pitch, chroma values and are selected using techniques like PCA and Redundant feature elimination technique to find the best feature , that is used to train the model. He explains about two approaches used in this research such as sequential model and single model approach for gender and age group prediction. The single model approach using Multi-Layer Perceptron gives better accuracy in order to predict the gender and age group of the speaker.

The author Evangelia Pantraki[2] addresses the problems faced in speech soft bio-metrics for predicting speakers age and gender and the author also feels that gender and age detection using face verification has more attention than speech verification, so he proposed a model to predict the gender and age of a person from speech using an framework named Parallel Factor Analysis 2 (PARAFAC2). The mentioned framework is applied in three metrics namely speech-utterance feature metrics, speaker age metrics and speaker gender metrics. The author concentrated more on the concept of vocal aging and age prediction than gender. But the author got an promising results with gender prediction model than age prediction model using PARAFAC2 framework, also mention reason for low prediction results of age model as that data-set used for age model is not sufficient as gender model.

In this paper[3], the author chin-chang-chen proposes an hierarchical method to predict the

gender and age from speech signal over an direct classification model. The proposed hierarchical model, firstly trained to predict the gender from the speech and the results are used to train the age predicting model. The gender predicting model predicts the gender as female and male, with this result the age predicting model predicts the age group as child, adult and elder. This paper contains an detailed explanation of the feature extraction process used in the hierarchical model, and the result of gender and age predicting model using hierarchical approach is compared with the direct classification model and shown an improved accuracy rate of 17.9% over the average. But though the model only got 50% accuracy ,which is considered as the drawback.

The author Y.C. Pan[4] made an early research in the field of speech recognition in the year 2007. The author's main focus is to improve the speech recognition method to recognize the verbal content of the speech with emotional hints, as the experiments shows that it is difficult to recover the verbal content when the speaker pronounces it in an emotional way. The proposed method can be used to recognize the speech based on the emotion expressed,that is an emotion recognition model is trained to identify the emotion of the speech and also speech recognition models are built to recognize the speech based on five class of emotions as neutral, happy, sad, anger and fear. Based on the prediction result from the emotion recognition model, the model for speech recognition is selected and used to recognize the verbal content of the speech.The drawback is the lowest accuracy attained from the trained models.

In this paper[5], the author Chen Jei mentioned that, the increasing demand for emotion recognition in many fields makes speech emotion recognition as an hot research topic in the speech field. So, the author proposed an speech emotion recognition model to predict the emotion based on its acoustic feature using Convolutional Neural Network. This paper contains an detailed explanation of feature extraction process and how the convolutional neural network architecture can be used to train emotion recognition model in order to get good prediction results, And in [6] , the author Likitha has proposed a speech emotion recognition model which predicts three class of emotions as happy, anger and sad based on its acoustic feature, the author mainly explains the feature extraction process, Mel Frequency Cepstral Coefficient method is used in the feature extraction process and achieved an 80% of accuracy.

In [7], the author Mahwish Pervaiz feels that variability in speech signal makes emotion recognition as a challenging tasks and the persodic or temporal feature and linguistic feature do not provide results with adequate accuracy. The author made two contribution through this paper,the first contribution is by proposing an two stage model for emotion recognition,

uses periodic and temporal features in the first stage and extract emotions from word segmentation combined with linguistic feature in second stage, as the second contribution the author proves through his experiments that considering age factor while training an emotion recognition model helps in improving the models accuracy.

The research work by Hao Hu [8] addresses speech emotion recognition as the challenging and important area of study in speech technology. The author introduces an model trained using spectral features of the speech and GMM supervector based SVM for predicting the emotion of the speech. And the author also proves that GMM Supervector based SVM works better than standard GMM. In [9], the author Norbert Braunschweiler proposes an Emotion Recognition in three ways, using only the speech to predict and using only the text to predict then concatenating the both models to predict the emotion of the speech, these model are trained using different deep learning architectures. The author also introduced an cross corpus fashion, where the trained models using one corpus will be tested using different corpora. The author mentions the data-sets used as corpora and used 6 different corpora to train and test the emotion recognition model.

2.2. OPEN PROBLEM IN EXISTING SYSTEMS

Over the years, there have been various attempts to develop models capable of accurately predicting gender, age, and emotions from speech. However, existing models used for predicting gender and age from speech based on acoustic features have shown a lack of accuracy, particularly with new inputs. This lack of accuracy means that these models heavily rely on the training data used, and they are not robust enough to generalize well to new data. Similarly, most emotion recognition models have relied on acoustic features more than linguistic features to predict emotions from speech. Although these models can predict emotions with good accuracy for three or four classes of emotions, they perform poorly when it comes to predicting five or six classes of emotions. This suggests that there are limitations to using acoustic features alone to capture the nuanced emotions present in speech. Additionally, no single system has been proposed to predict gender, age, and emotion from speech simultaneously. Instead, separate models have been developed to predict each of these features. While separate models for each feature are useful, it is likely that these features are interdependent, and predicting them together would yield more accurate results. Therefore, there is a need for more comprehensive models that can simultaneously predict gender, age, and emotion from speech while taking into account the complex relationships between these features. The issue with predicting age from speech is that it can be challenging to differentiate between individuals of different ages, especially for those who are in the same age group. As such, age prediction models may not always be

accurate. Additionally, predicting gender from speech is often limited by societal norms and biases that may affect the way individuals speak. For example, some people may speak in ways that do not conform to traditional gender expectations. Furthermore, the use of only acoustic features to predict emotions from speech can be limited, as emotions can also be conveyed through linguistic features such as word choice and sentence structure. This limitation can be addressed by developing more sophisticated models that can capture both the acoustic and linguistic features of speech. In conclusion, predicting gender, age, and emotion from speech is a complex task that is still being improved upon. While there are currently separate models for each of these features, there is a need for more comprehensive models that can predict all three features simultaneously. Developing such models will require further research and innovation in this field.

CHAPTER 3

3.1. FEASIBILITY STUDIES / RISK ANALYSIS OF THE PROJECT

Training models to predict gender, age, and emotion using speech features has the potential to be a valuable tool in a wide range of applications, from speech recognition and transcription to marketing and advertising. However, there are also a number of risks associated with these models that must be taken into account.

One of the main risks associated with training models to predict gender, age, and emotion is the potential for bias. These models may be trained on datasets that are not representative of the population as a whole, leading to inaccurate or skewed predictions. For example, a model trained on a datasets that over-represents young, male voices may have difficulty accurately predicting the gender, age, and emotions of older, female speakers. Similarly, a model trained on a datasets of speakers from a particular region or language may not be applicable to speakers from other regions or languages. Another risk associated with these models is the potential for privacy violations. Speech contains a wealth of personal information, including gender, age, and emotional state, as well as potentially sensitive information about a person's identity, such as their name or occupation. If these models are used in contexts where privacy is important, such as in healthcare or legal settings, there is a risk that this information could be accessed or shared without a person's consent.

In addition, there is a risk that these models could be used in ways that are discriminatory or harmful. For example, if a model is used to screen job applicants based on their age or gender, this could lead to discrimination against older or female candidates. Similarly, if a model is used to target advertisements based on a person's emotional state, this could be seen as manipulative or exploitative. There are also technical risks associated with training these models. Acoustic and lexical features of speech can be highly complex, and it can be difficult to ensure that models are accurately capturing all relevant information. Additionally, the quality and availability of training data can be a challenge, particularly for models that require large, diverse datasets. Finally, there is always a risk of overfitting, where a model becomes too closely tailored to the training data and is unable to generalize to new data. To mitigate these risks, it is important to approach the development and use of these models with care and caution. This may involve ensuring that training data is representative

of the population as a whole, implementing privacy protections and consent mechanisms where appropriate, and carefully considering the potential ethical implications of using these models in different contexts. It may also involve ongoing monitoring and evaluation of these models to ensure that they are performing accurately and ethically. By taking these steps, we can help ensure that these models are used.

3.2. SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

To develop a speech analysis system that can predict the age, gender, and emotion of a speaker using machine learning and deep learning algorithms , you will need the following software requirements:

1. Programming Language:

we will need a programming language such as Python to develop the system. Python have libraries and packages for speech processing and machine learning. Python is a high-level, interpreted programming language that was first released in 1991 by Guido van Rossum. It is designed to be easy to read, write and understand. Python is known for its clear syntax, dynamic typing and its ability to be used in many different domains such as web development, scientific computing, data analysis, machine learning, and artificial intelligence. Python has a large standard library and also supports a vast array of third-party packages and libraries, including NumPy, Pandas, Matplotlib, Scikit-learn, and TensorFlow, which makes it a powerful tool for machine learning and data analysis. Python is an interpreted language, meaning that code can be executed directly without being compiled first, which makes it highly accessible and easy to use for beginners. It is also cross-platform, meaning that it can be used on different operating systems, including Windows, Linux, and macOS. Python is widely used in the development of speech analysis systems, as it provides many powerful libraries for processing and analyzing speech data, such as SpeechRecognition, PyAudio, and Praat. Additionally, Python's machine learning libraries, such as Scikit-learn and TensorFlow, are frequently used to develop models for predicting age, gender, and emotion based on speech features.

2. Speech Processing Libraries:

We will need libraries such as PyAudio, librosa, or Speech_Recognition to extract acoustic features from speech. These libraries allow you to record, read, and process audio files. Speech processing libraries are essential tools in the development of speech analysis

systems. These libraries provide functionalities to record, read, and process audio files for speech processing. Some popular speech processing libraries include PyAudio, librosa, and Speech_Recognition. PyAudio is a cross-platform audio I/O library that allows audio streaming and audio recording. It can be used to record audio files from a microphone, read audio files from disk, and play audio files. Librosa is a Python library for analyzing audio signals and music. It provides tools for computing various audio features such as the mel-frequency cepstral coefficients (MFCC), spectrograms, and chroma features. Speech_Recognition is a Python library that supports various speech recognition engines, including Google Cloud Speech, PocketSphinx, and Wit.ai. It allows you to convert speech to text and perform other speech recognition tasks. These speech processing libraries are critical in extracting the acoustic features from speech that are needed to predict age, gender, and emotion. These features can include pitch, loudness, speaking rate, spectral characteristics, and prosody. The libraries provide a wide range of functionalities that can be customized and combined to develop a speech analysis system that meets specific needs. With these libraries, developers can easily implement the audio processing and feature extraction parts of a speech analysis system, without having to develop everything from scratch. Overall, these speech processing libraries simplify the development process of speech analysis systems by providing the necessary tools for audio processing and feature extraction. This allows developers to focus on building the machine learning models that predict age, gender, and emotion based on these features. By leveraging these libraries, developers can build powerful and accurate speech analysis systems that can be used in various fields, including healthcare, education, entertainment, and security.

3. Machine Learning Libraries:

We will need machine learning libraries such as scikit-learn, TensorFlow, or Keras to train and test your models. These libraries provide various algorithms for classification, regression, and clustering. Machine learning libraries are crucial components when developing a speech analysis system to predict age, gender, and emotion. They enable you to implement different machine learning models and train them with your data. Some popular libraries for machine learning include scikit-learn, TensorFlow, and Keras. Scikit-learn is an open-source library that provides simple and efficient tools for data analysis and machine learning. It includes a wide range of classification, regression, and clustering algorithms, and it supports both supervised and unsupervised learning. Scikit-learn also includes preprocessing tools for feature extraction, dimensionality reduction, and data normalization. TensorFlow is a popular open-source library developed by Google for building and training machine learning models. It provides a flexible and efficient platform for implementing and deploying machine learning algorithms across a range of devices and

platforms. TensorFlow supports a variety of neural network architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep neural networks (DNNs). Keras is a high-level neural network API written in Python that is built on top of TensorFlow. It provides a simple and intuitive interface for designing and training deep learning models. Keras supports a wide range of neural network architectures, including CNNs, RNNs, and DNNs, and it includes several pre-trained models that can be fine-tuned for specific applications. These machine learning libraries enable you to train and test different models using your data, and you can choose the most appropriate algorithm for your specific application. Additionally, these libraries allow you to optimize your models by tuning hyperparameters and evaluating performance metrics.

4. Natural Language Processing Libraries:

We will need a library named as NLTK to extract lexical features from speech. This library allow you to process and analyze text data. Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence that deals with the interaction between computers and human language. It involves processing, analyzing, and understanding human language in a way that enables computers to interpret and respond to it. One of the most popular NLP libraries is the Natural Language Toolkit (NLTK), which is a platform for building Python programs to work with human language data. NLTK provides various tools for text processing and analysis, such as tokenization, stemming, part-of-speech tagging, and named entity recognition. These tools can be used to extract lexical features from speech, such as the frequency and distribution of certain words, phrases, or linguistic patterns. These features can then be used to train machine learning models to predict age, gender, or emotion. For example, NLTK can be used to preprocess and analyze a transcript of a speech to extract features such as the number of pronouns used, the length of sentences, or the frequency of certain words related to emotions. These features can then be used as inputs to a machine learning algorithm to predict the speaker's age, gender, or emotional state. NLTK also provides access to various datasets for language analysis, such as the Brown Corpus, which is a collection of text samples from different genres and registers of English. These datasets can be used to train and test machine learning models for speech analysis. In summary, NLTK is a powerful tool for extracting lexical features from speech, which can be used in combination with acoustic features to develop machine learning models for speech analysis.

5. Data Analysis and Visualization Tools:

We will need tools such as Pandas, NumPy, or Matplotlib to analyze and visualize the data. These tools provide various functions for data analysis, manipulation, and visualization. Data analysis and visualization tools are essential components of any data science project, including speech analysis systems. Pandas is a powerful data analysis library that provides data structures for efficient data manipulation and analysis. It also provides data cleaning and preprocessing tools that help in data cleaning and handling missing values. NumPy is a numerical computing library that is used for numerical operations and mathematical computations. It provides an array data structure for efficient data storage and manipulation. Visualization is a crucial part of data analysis as it helps to present the results in a clear and concise manner. Matplotlib is a popular data visualization library in Python that provides various functions for creating 2D and 3D plots. It supports various types of plots such as line, scatter, bar, histogram, and many more. It is also used for creating visualizations for time series, heat maps, and network graphs. Together, these tools provide a comprehensive data analysis and visualization environment for speech analysis systems. They allow developers to analyze the data, visualize the results, and gain insights from the data. This is particularly useful for interpreting the results of the system and making informed decisions based on the data.

6. Integrated Development Environment (IDE):

We will need an IDE such as Jupyter Notebook to write, debug, and execute your code. This IDE provides a user-friendly interface for writing and executing code. An Integrated Development Environment (IDE) is a software application that provides a comprehensive environment for writing, debugging, and executing code. IDEs offer a user-friendly interface for developers to write and manage code in a single application. For our speech analysis system, we will need an IDE that supports Python programming language and allows us to interact with various libraries and packages we use in our project. One such popular IDE is Jupyter Notebook. It is a web-based interactive computing environment that allows you to create and share documents containing live code, equations, visualizations, and narrative text. Jupyter Notebook supports a variety of programming languages, including Python, R, and Julia, making it an ideal choice for data science and machine learning projects. Jupyter Notebook provides a user-friendly interface for writing, executing, and debugging code. It allows you to run code in individual cells, making it easier to test and debug your code. You can also add markdown text to your notebook to provide explanations, instructions, or comments for your code. This makes it easier to document your code and share it with others. Jupyter Notebook also allows you to install and use various libraries and packages we need in our speech analysis system. It supports popular data science and machine learning

libraries such as NumPy, Pandas, Scikit-Learn, TensorFlow, and Keras. You can install these libraries directly from the notebook using pip, the Python package manager. Overall, Jupyter Notebook provides an excellent platform for developing and testing our speech analysis system. It supports all the tools and libraries we need and provides a user-friendly interface for writing, executing, and documenting our code.

Overall, developing a speech analysis system to predict age, gender, and emotion of a speaker using machine learning algorithms requires a combination of programming languages, libraries, tools, and platforms. You can choose the software requirements based on your project requirements, budget, and expertise.

3.3. SYSTEM USE CASE:

A use case diagram for a speech analysis system that predicts age, gender, and emotion of the speaker would typically consist of several actors and use cases. The actors in the system are the individuals or systems that interact with the speech analysis system. One of the primary actors in the system would be the user, who interacts with the system through a user interface to provide input, such as an audio file or a live recording. Another actor might be a speech recognition engine, which is responsible for converting speech signals into digital data that can be analyzed. The main use cases in the system would be related to the analysis of the speech signal to predict the age, gender, and emotion of the speaker. These use cases might include:

- Speech input: This use case represents the user providing an audio file or live recording as input to the system.
- Speech recognition: This use case represents the speech recognition engine converting the speech signal into digital data that can be analyzed.
- Acoustic analysis: This use case represents the analysis of the acoustic features of the speech signal, such as pitch, tone, and cadence, to predict the age and gender of the speaker.
- Lexical analysis: This use case represents the analysis of the lexical features of the speech signal, such as the choice of words and the context in which they are used, to predict the emotion of the speaker.

- Output generation: This use case represents the generation of a report or output that summarizes the age, gender, and emotion of the speaker based on the analysis of the speech signal.

Other use cases might include user authentication, system maintenance, and error handling. The use case diagram would typically show the relationships between the actors and the use cases in the system. For example, the speech recognition engine would be associated with the speech input and speech recognition use cases, while the acoustic and lexical analysis use cases would be associated with the output generation use case. The user would be associated with all the use cases, as they are the primary user of the system. Overall, the use case diagram provides a high-level overview of the speech analysis system and the interactions between the various actors and use cases. It helps to ensure that all the requirements of the system are identified and that the system is designed to meet the needs of the user.

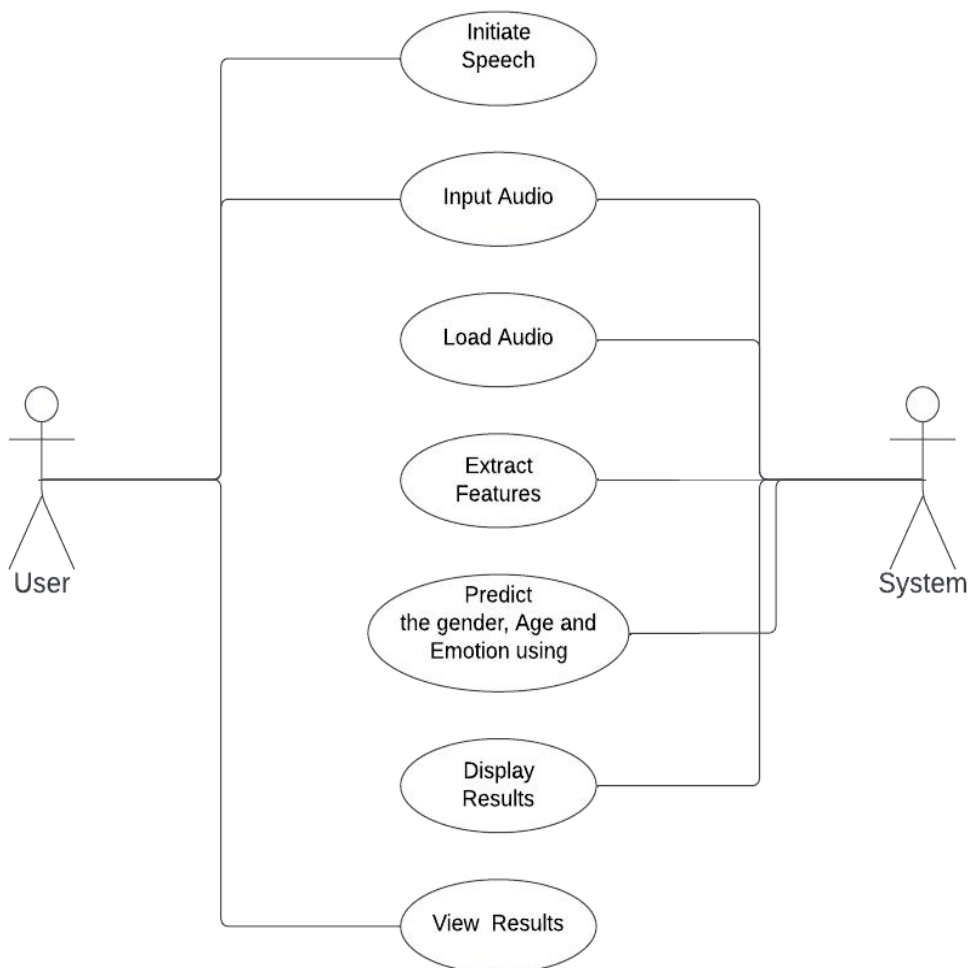


Fig 3.1. UseCase Diagram of Speech Analysis System

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

4.1. SELECTED METHODOLOGY OR PROCESS MODEL

- The proposed Speech Analysis System is developed to predict gender, Age and Emotion of the speaker using both acoustic and lexical feature. The models to predict the gender, age and emotion of the speaker are developed with large speech and text samples and using Machine Learning and Deep Learning Architecture.
- The first step is to collect the appropriate data-sets that can be used to train the models to predict the gender, age and emotion of the speaker. The gender and age model are trained using same data-set which contains speech samples but the emotion model using acoustic feature used separate data-set with speech sample and emotion model using Lexical feature used data-set with text samples.
- The next step is data pre-processing, where the selected Speech and text data-sets are analyzed to remove null and inappropriate values. And the speech samples are stacked in an new data-frame as if to make it easier for the next step.
- After pre-processing the data, the acoustic characteristics from the speech samples are extracted using various feature extraction methods and are stored into an data-frame, which will be used in the to train the models that predicts the Gender, Age and Emotion of the speaker using the acoustic characteristics.
- Whereas in the emotion recognition model using lexical feature, the preprocessed text data-set undergoes normalization process. Where many functions are libraries are used to remove stopwords, punctuation, numbers and small sentences and all the letters present in the dataset are reduced to lower case letters and it gets stored in an new data-frame, which will be used to train the model to predict the emotion of the speaker using the lexical content.
- After the normalization and feature extraction part, the data-sets created newly will used to train the models to predict gender ,age and emotion of the speaker. The model to predict the age and gender are trained with machine learning algorithms and the best

one will be selected for further process.

- Where as the model to predict the emotion of the speaker using acoustic features are trained using Deep learning architecture, as we are using large number speech samples to train the Speech Emotion Recognition model.
- The model to predict the emotion of the speaker using the lexical feature like the content or word expressed by the speaker is named as Text Emotion Recognition model. It is also trained using Machine learning algorithms and the best performed one will be selected to predict the new samples.
- To predict the new sample with the trained model, we first extract the features as we did in the training process, the extracted feature will then be used predict the speakers gender, age and Emotion.
- As the text emotion recognition model are trained using text data-set, it only takes text sample as the input to predict, but if we want to predict the emotion by giving an speech sample, we have to convert the speech sample to text using speech recognition methods.

4.2. ARCHITECTURE / OVERALL DESIGN OF PROPOSEDSYSTEM

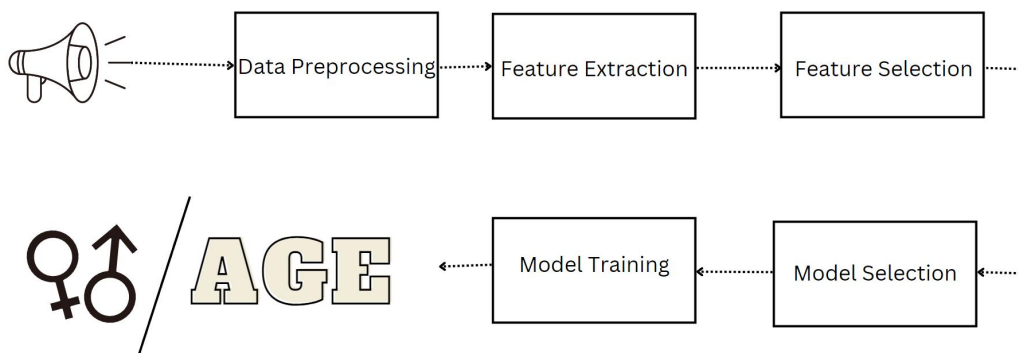


Fig 4.1 Flowchart of Gender and Age Model

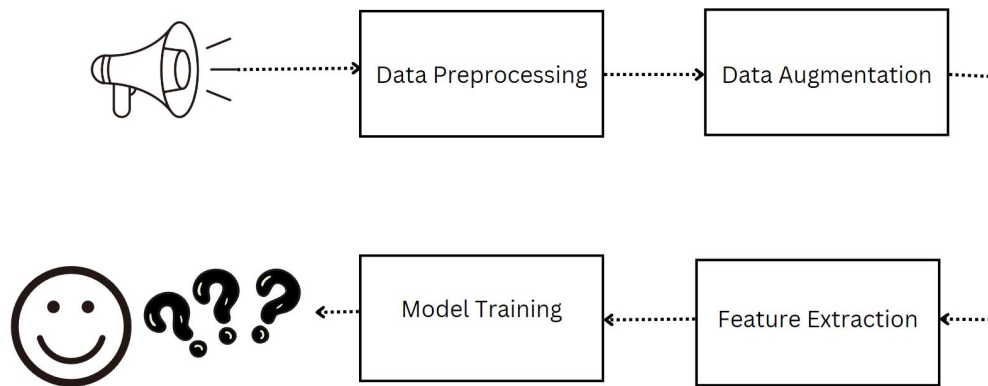


Fig 4.2 Flowchart of Speech Emotion Model

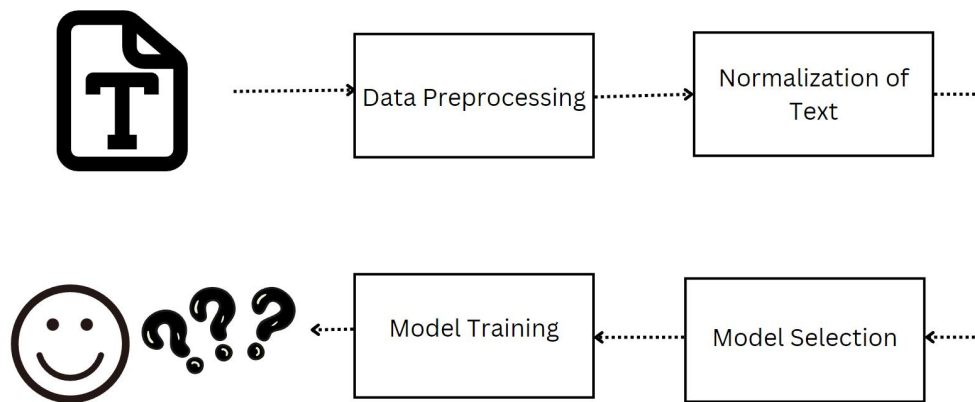


Fig 4.3 Flowchart of Text Emotion Model

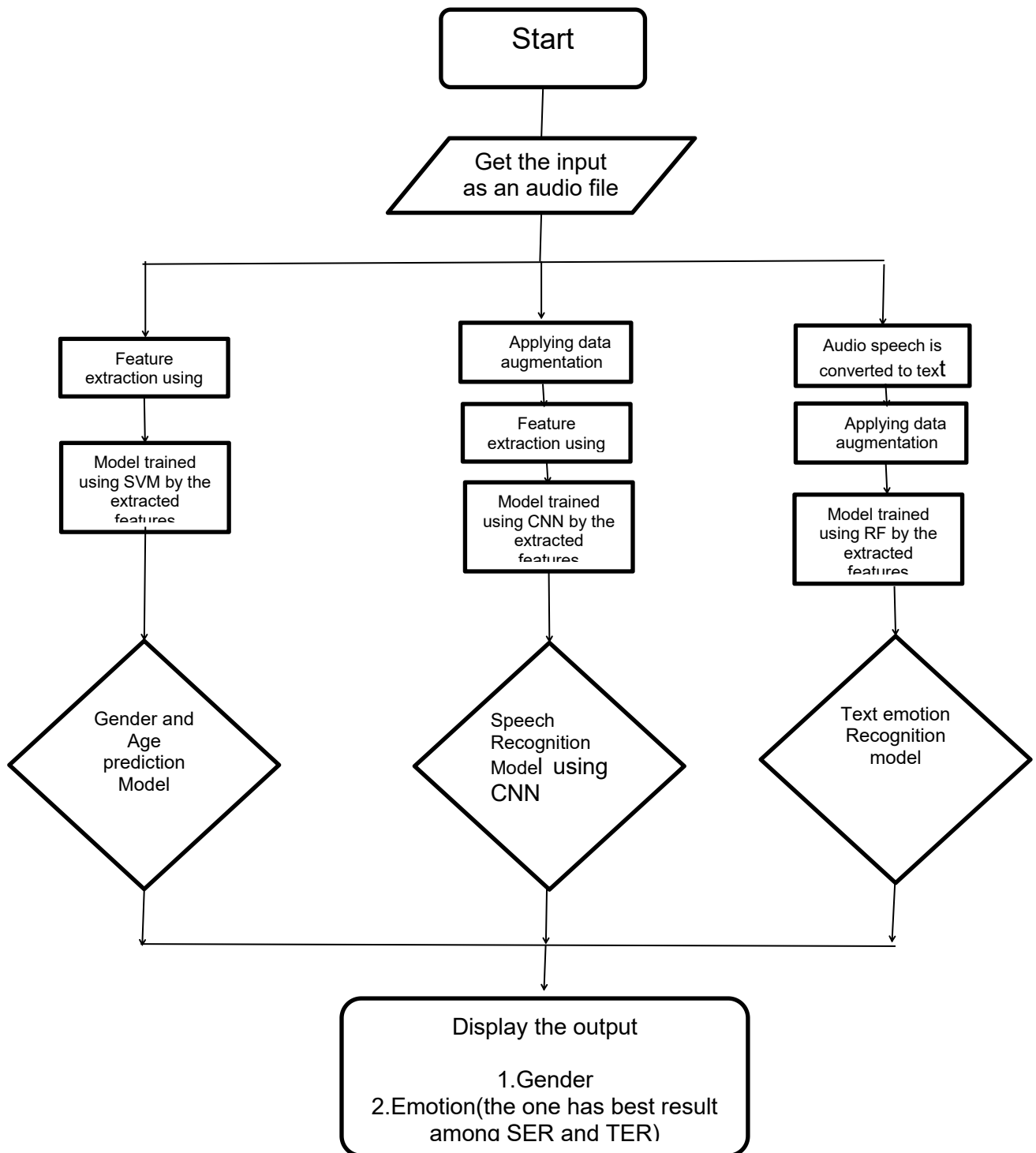


Fig.4.4 Flowchart of the Speech Analysis System

4.3. DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION AND TESTING PLAN OF THE PROPOSED SYSTEM

- **PYTHON:** Python is a high-level programming language with a simple syntax that emphasizes readability and ease of use. It is an interpreted language, meaning that code is executed directly without the need for compilation. Python is widely used for a variety of purposes, including web development, data analysis, machine learning, and automation. Its extensive library of modules and packages makes it a popular choice for developers, and its open-source nature has led to a vibrant and supportive community.
- **SPEECH RECOGNITION LIBRARIES:**
 - **Pyaudio** is a cross-platform audio I/O library for Python that provides easy access to audio input and output devices. It can be used to record and play audio from different sources.
 - **Librosa** is a Python library for analyzing and processing audio signals. It provides tools for loading and manipulating audio files, computing spectrograms and other time-frequency representations, extracting features from audio signals, and more. It is commonly used for speech and music signal processing applications.
- **MACHINE LEARNING LIBRARIES:**
 - **Scikit-learn** is a popular open-source machine learning library for Python that provides tools for data preprocessing, model selection, and evaluation, as well as a variety of supervised and unsupervised learning algorithms.
 - **TensorFlow** is an open-source platform for building and deploying machine learning models developed by Google. It provides a range of tools, including a flexible computational graph system, neural network models, and algorithms for training and inference.
 - **Keras** is an open-source high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It provides a user-friendly interface for building and training neural networks, including various pre-built models and layers.

- **NATURAL LANGUAGE PROCESSING:** NLP (Natural Language Processing) is a subfield of AI that focuses on enabling machines to understand, interpret, and generate human language. It involves applying algorithms, statistical models, and linguistic rules to analyze and manipulate natural language data, such as text, speech, and images. NLP has various applications, including sentiment analysis, machine translation, text summarization, chatbots, and voice assistants.

- **DATA ANALYSIS AND DATA VISUALIZATION TOOLS:**
 - **NumPy** is a Python library used for working with large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions.

 - **Pandas** is a library built on top of NumPy and provides easy-to-use data structures and data analysis tools for handling tabular data, including reading and writing various file formats, merging, filtering, and grouping data.

 - **Matplotlib** is a plotting library used to create static, animated, and interactive visualizations in Python. It provides a wide range of options for creating plots, histograms, bar charts, scatterplots, and more.

- **INTEGRATED DEVELOPMENT ENVIRONMENT:** Jupyter Notebook is an open-source web application that allows users to create and share documents containing live code, equations, visualizations, and narrative text. It supports various programming languages
 - including Python, R, Julia, and many others. It offers an interactive environment for data analysis, prototyping, and exploration, making it a popular tool among data scientists, researchers, and educators. Jupyter Notebook provides an easy-to-use interface that promotes reproducible and collaborative research.

4.4. PROJECT MANAGEMENT PLAN

- **Project Objective:** The objective of this project is to develop a speech analysis system that can predict the age, gender, and emotion of a speaker using acoustic features and emotion using lexical features.
- **Project Scope:** The scope of the project includes collecting speech data from different sources, preprocessing and cleaning the data, extracting relevant acoustic and lexical features, developing machine learning models to predict age, gender, and emotion, and evaluating the performance of the models.
- **Project Deliverables:** The key deliverables of the project include a speech dataset, preprocessed and cleaned data, feature extraction code, machine learning models, and a report summarizing the project outcomes.
- **Project Resources:** The resources required for the project include a team of data scientists and machine learning experts, access to speech data sources, computing resources, and relevant software and libraries for data processing, feature extraction, and model development.
- **Project Risks and Mitigation:** Risks associated with the project include data quality issues, computational limitations, and model performance challenges. These risks can be mitigated through careful data selection and preprocessing, efficient use of computing resources, and rigorous model evaluation techniques.

CHAPTER 5

IMPLEMENTATION

5.1. DEVELOPMENT AND DEPLOYMENT SETUP:

Developing and Deploying the Speech analysis system involves several steps and the common steps involves in each Gender and Age predicting model, Speech Emotion Recognition and Text Emotion Recognition models are explained,

Data Collection:

Collect a large dataset of speech recordings with labeled emotion classes. This dataset should represent a diverse range of emotions and be balanced in terms of the number of samples for each emotion class. And also text samples with labeled emotion class to train model using lexical features.

Data Preprocessing:

In this step, the selected Speech and text data-sets are analyzed to remove null and inappropriate values. And the speech samples are stacked in an new data-frame as if to make it easier for the next step.

Feature extraction and Normalization:

- In the gender, age and emotion recognition model using acoustic characters, the acoustic characteristics from the speech samples are extracted using various feature extraction methods and are stored into an data-frame, which will be used in the to train the models that predicts the Gender, Age and Emotion of the speaker using the acoustic characteristics.
- Whereas in the emotion recognition model using lexical feature, the preprocessed text data-set undergoes normalization process. Where many functions are libraries are used to remove stopwords, punctuation, numbers and small sentences and all the letters present in the dataset are reduced to lower case letters and it gets stored in an new data-frame, which will be used to train the model to predict the emotion of the speaker using the lexical content.

Model Training:

After the normalization and feature extraction part, the data-sets created newly will be used to train the models to predict gender, age and emotion of the speaker. The model to predict the age and gender are trained with machine learning algorithms and the best one will be selected for further process. Whereas the model to predict the emotion of the speaker using acoustic features are trained using Deep learning architecture, as we are using large number speech samples to train the Speech Emotion Recognition model. The model to predict the emotion of the speaker using the lexical feature like the content or word expressed by the speaker is named as Text Emotion Recognition model. It is also trained using Machine learning algorithms and the best performed one will be selected to predict the new samples.

Model Evaluation:

Evaluate the trained model on the test set to estimate its performance on new and unseen data. The evaluation metrics used can be accuracy, precision, recall, F1-score, and AUC-ROC.

5.2. MODULES IMPLEMENTED

The proposed Speech Analysis System to identify Gender, age of the speaker using acoustic features and emotion of the speaker using both acoustic and lexical characters are developed as three modules which are Gender and Age recognition Model, Speech Emotion Recognition model and Text Emotion Recognition model.

● GENDER AND AGE RECOGNITION MODELS

To predict the gender and age of the speaker using the acoustic feature, we have trained to separate models with same methodology and same dataset. Some of the main steps involved in the process are explained in detail.

➤ DATASET:

Very few datasets are openly available with speech samples and their related gender and age in-formations. In this proposed system we have used Mozilla Common Voice dataset, which contains large number of speech samples, but only a small number of speech samples contains their related demographic features like gender, age, accent and . We managed to extract only those samples that contains the aforementioned demographic features and applied in our model to predict age and gender. Fig.5.1 explains the available

samples of age groups with their gender information.

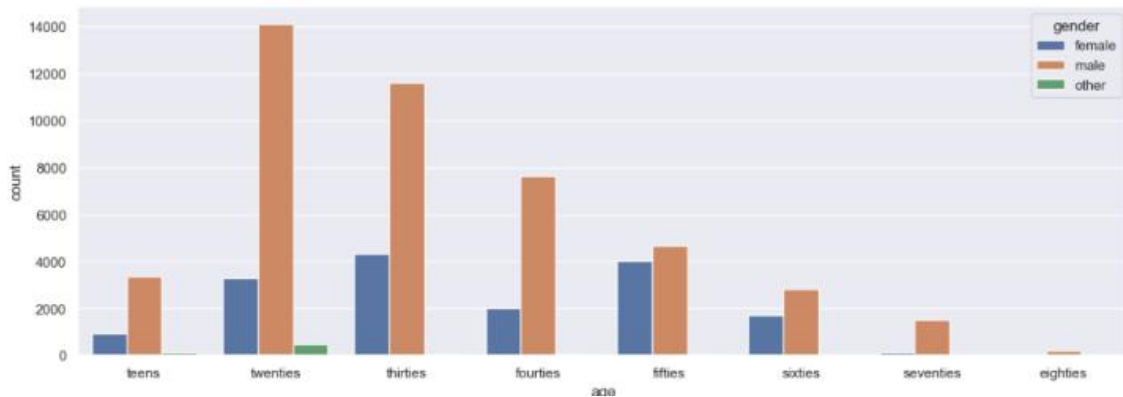


Fig.5.1. Statistical Representation of Gender and Age Dataset

➤ **FEATURE EXTRACTION:**

Feature extraction is the process of extracting the features from anything like images or audio signals which then will be used to train the machine learning model to predict the desired results. Here we extract the acoustic characteristics from the speech signals to train our ML model to predict gender and age of the speaker. We extract four types of features from the speech signal with 22 parameters to train our model. The types features extracted from speech signals are explained below,

- ◆ **Spectral Centroid:** It indicated the center of mass of the frequencies distributed in the audio signal.
- ◆ **Spectral Bandwidth:** It is derived after the spectral centroid that is the range of frequency in the audio signal around the centroid.
- ◆ **Spectral Rolloff:** It is the point where a specified percentage of total spectral energy lies.
- ◆ **MFCC:** Mel Frequency Cepstral Coefficient(MFCC) are a set features that are extracted and used most commonly in the field of speech recognition, these features from MFCC are derived from the Mel Frequency cepstrum, by taking logarithm of the power spectrum and transforming to the Mel frequency scale and then applying DCT to obtain cepstral coefficient.

➤ **MODEL TRAINING**

The extracted features using aforementioned methods along with the gender information are used to train the model to predict the age of the speaker, same as the model to predict the gender, uses the extracted features and the age group information to train the model. As explained above, the models to predict age and gender are developed separately using the same extracted feature but as an extra feature the age group is considered for gender recognition model and gender information is considered for age group recognition. The performance of Support Vector Machine and Random Forest on the extracted features are evaluated. SVM performs well than Random Forest, so SVM is used to train the models to predict age and gender.

● **SPEECH EMOTION RECOGNITION:**

Recognizing the emotion from the speech signal i.e., using the acoustic characteristics is an complicated task, here in this paper we propose a method to recognize the emotion by including data augmentation process and we tried to extract Eight class of emotions such as Anger, Sad, Fear, Happy, Disgust, Surprise, Calm and Neutral. fig.4.2 explains the process of Speech Emotion Recognition model.

➤ **DATASET**

Many labeled datasets are available openly with speech samples and their related emotional details expressed in the speech. The proposed Speech Emotion Recognition model is trained using four popular datasets, named as RAVDESS, SAVEE, TESS and CREMA-D. RAVDESS contains eight emotion class (Anger, Sad, Fear, Happy, Disgust, Surprise, Calm and Neutral) of speech samples, SAVEE contains Seven emotion class (Anger, Sad, Fear, Happy, Disgust, Surprise, and Neutral) of speech samples, TESS and CREMA-D contains Six emotion class (Anger, Sad, Fear, Happy, Disgust, and Neutral) of speech samples, these speech samples are combined together and stored in a new data-frame which will then be used to extract the features. The fig.5.2 shows the available speech samples of different emotions from the combined data-frame.

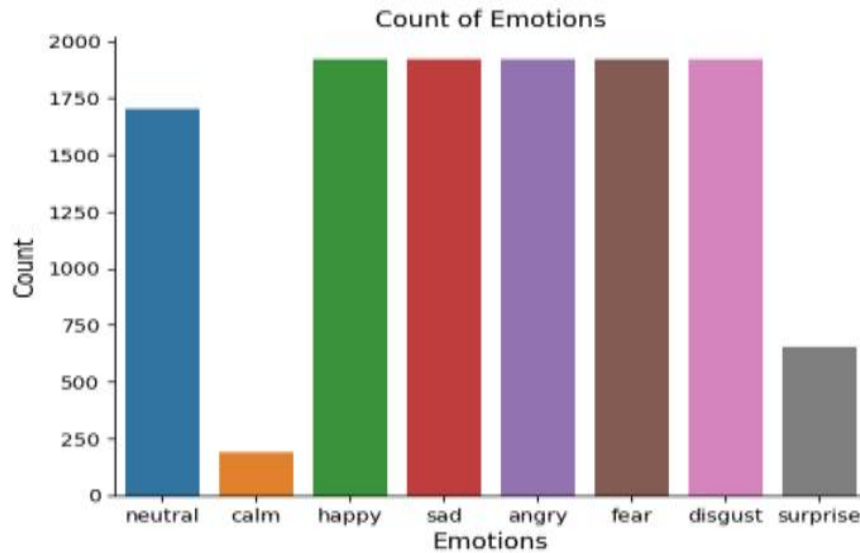


Fig.5.2. Statistical Representation of Speech Emotion Dataset

➤ DATA AUGMENTATION

Data Augmentation is commonly used in machine learning to improve the training dataset in order to prevent the model from over-fitting. In audio processing, data augmentation is applied to audio samples using various transformation techniques to create new samples which then be used to train the model. Here in this proposed model, we used four augmentation techniques to transform the training data, those techniques are mentioned below,

- ◆ **Noise:** it simply adds random noise to audio signal by calculating its maximum amplitude value then multiplying it with some random value, gives the noise amplitude which then be added to the audio signal.
- ◆ **Stretch:** It is used to stretch or compress the audio signal, here we compress the audio signal to 20%.
- ◆ **Shift:** transforms the audio signal by shifting certain number of samples to left or right.
- ◆ **Pitch:** The pitch of the audio signal is shifted using certain factor, here we shift down the pitch by 30%.

➤ FEATURE EXTRACTION

As mentioned above in the gender recognition model, feature extraction is the process extracting the acoustic features from the speech signals. Here in this Speech Emotion Recognition model those features are used to train the DL model that predicts the emotion of the speaker. As we are using Deep Learning architecture, after augmenting the training dataset, five types of features are extracted from the speech signal with 161 parameters to train our model. The types of features extracted for Speech Emotion Recognition model are explained below,

- ◆ **Zero Crossing Rate:** ZCR is used to estimate the temporal properties of audio signal such as their rhythm or tempo and also this feature is used to differentiate the voiced and unvoiced segments of speech.
- ◆ **Chroma STFT:** This feature is calculated using the Short Time Fourier Transform(STFT) of the audio signal and used to identify tonal characteristics of speech signal.
- ◆ **Mel Frequency Cepstral Coefficient:** As mentioned in the gender recognition model, MFCC is the most commonly used feature extraction technique in speech recognition and used to capture the timbral characteristics of speech signal.
- ◆ **Root Mean Square:** It is used to measure the total energy present in the audio signal by taking square roots of the mean of the squared values of the audio sample. It is also an most commonly used method, which provides the information about the loudness of the audio signal.
- ◆ **Mel Spectrogram:** It is used to capture the distribution of energy in the audio signal across different frequency bands, which then can be used to train the ML model.

The above mentioned feature extraction methods are used to extract the feature from the speech samples.

➤ MODEL TRAINING

After extracting the feature from the speech sample, we used those features to train our Deep Learning Model. As large number of features are extracted from the speech samples, CNN architecture is used to train our model. Usually Convolutional Neural Network requires large dataset to train and are often used in Image processing but as mentioned in

[6], CNN can also be effective in Audio Processing. So we used an CNN architecture to train our model that can predict Eight class of emotions such as Anger, Sad, Fear, Happy, Disgust, Surprise, Calm and Neutral.

● TEXT EMOTION RECOGNITION:

This model is trained to predict the emotion of the speaker by considering the lexical features i.e, the words or content expressed by the speaker. This model is trained using Natural Language Processing Techniques with text dataset to predict Six class of emotions such as Anger, Sad, Joy, Disgust, Surprise and Fear.

➤ DATASET

We have used an open sourced NLP dataset that contains Large number of text sentences and their related emotion as their label. This dataset is separately available for Training, Testing and Validation parts and are specially built and mentioned be used in an model to classify the emotion using the textual content. Fig.5.3 shows the available data in training dataset.

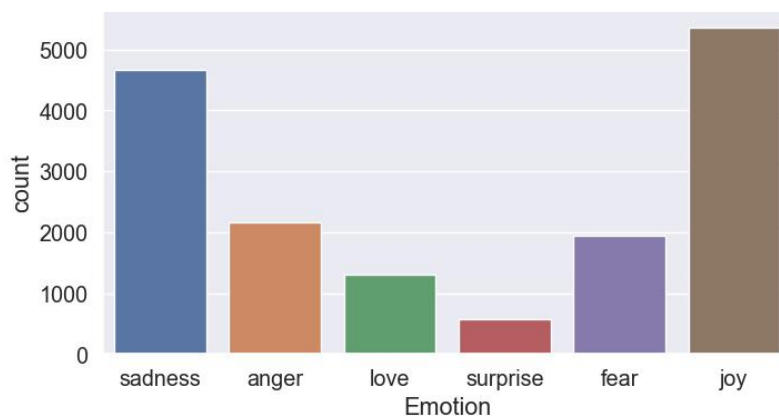


Fig.5.3. Statistical Representation Text Emotion Dataset

➤ NORMALIZING TEXT

In this process we tried to remove the unwanted informations from the data, which will be used for training and testing the model. We use functions to lemmatize words, converts all letters to smaller case and to remove numbers, punctuation, URLs, small sentences and stopwords, stopwords are some commonly used words like 'a' , 'an', 'the' 'I' etc. , are removed using a list of predefined stopwords. Applying all these function to remove and convert the words makes the texts more easier for the machine to understand and learn

to train the model. This process is mentioned as normalization of the text. We apply this to our dataset before training the model and apply this when we test the model with new data.

➤ **MODEL TRAINING**

After normalizing the text data, we apply that to train our model. We have tried to evaluate the performance of four model that are trained using Machine Learning algorithms like Logistic Regression, Decision tree, Support Vector Machine and Random Forest, to find the best model with better results. Among the four trained model, Random Forest has the best results and are used in the final model to predict the six class of emotions such as Anger, Sad, Fear, Disgust, Surprise and Joy from the textual content expressed by the speaker in the speech.

5.3. ALGORITHMS

We have used multiple algorithms to check performance of each with our Gender, age and emotion prediction models. The better performed algorithms that we have used in our resulting models are explain in brief below,

● **SVM IN GENDER EMOTION RECOGNITION MODEL :**

The model is an instance of the Support Vector Classification (SVC) algorithm, which is a type of supervised learning algorithm used for classification tasks. The parameter "C" controls the trade-off between maximizing the margin and minimizing the misclassification of training examples. The higher the value of C, the more the model will prioritize correctly classifying training examples, potentially leading to overfitting. The parameter "gamma" determines the shape of the decision boundary. A lower value of gamma will result in a decision boundary that is smoother and more generalized, while a higher value will result in a more complex and overfitted decision boundary. The 'scale' value for gamma is used to automatically scale the input features.

● **CNN IN SPEECH EMOTION RECOGNITION MODEL:**

This is a Convolutional Neural Network (CNN) algorithm for speech analysis system. The model is created using the Sequential class from the Keras library. The first layer of the model is a 1D convolutional layer with 256 filters of size 8 and padding set to 'same'. The input shape is specified as the shape of the training data, with the number of

columns representing the number of features.

An activation function, Rectified Linear Unit (ReLU), is added after the convolutional layer to introduce non-linearity into the model. Another convolutional layer with the same filter size, padding and number of filters is added along with a Batch Normalization layer to normalize the activations of the previous layer.

A Dropout layer is added to reduce the overfitting of the model by randomly dropping some neurons during training. A MaxPooling1D layer is added to reduce the spatial dimensionality of the output. These layers are then repeated with smaller number of filters in the following layers.

The output of the convolutional layers is then flattened and passed through a fully connected layer with the number of nodes equal to the number of target classes. The output of this layer is passed through a Softmax activation function to output probabilities for each class.

The model is compiled with the Adam optimizer and categorical cross-entropy loss function. The model is trained on the training data and evaluated on the test data. The summary function is used to print the details of the model including the number of parameters and the output shape of each layer.

● **RANDOM FOREST IN TEXT EMOTION RECOGNITION MODEL:**

The RF algorithm stands for Random Forest, which is a type of ensemble learning algorithm for classification, regression and other tasks that operates by constructing multiple decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It randomly selects a subset of features and samples to train each tree, which reduces overfitting and improves generalization. The "random_state = 0" parameter is used to set the seed of the random number generator, which ensures reproducibility of the results. The "train_model" function is used to train the Random Forest classifier on the X_train and y_train data. Once trained, the RF model can be used to predict the class labels of new, unseen data.

CHAPTER 6

RESULTS AND DISCUSSION

The proposed Speech Analysis System is a comprehensive tool that uses advanced machine learning algorithms to accurately predict various features from speech and text samples. The system is developed using Python and its main objective is to predict the speaker's age, gender, and emotion. To achieve this goal, the system extracts acoustic and lexical features from the speech and text samples, respectively. The dataset named as Mozilla Common Voice dataset is used to train the gender and age recognition model, contains male, female and age wise categorized samples like teens, twenties, thirties, forties, fifties, sixties, seventies and eighties. This dataset contains more number of male speech samples than female and more number of age speech samples as teens, twenties, thirties, forties, fifties, sixties than seventies and eighties. The gender recognition model is trained by extracting the acoustic features from the speech samples and using the Support Vector Machine (SVM) algorithm and achieved an accuracy of 86%, while testing with new data the model predicts the gender of male with higher accuracy than female. As female speech samples are less in the used dataset the model is trained relying more on male speech samples, this results in the difficulty for the model to predict female. Similarly, the age prediction model is also trained using acoustic features and the SVM algorithm, achieving an accuracy of 46%, which is quite good compared to existing models. The age groups predicted by the model include teens, twenties, thirties, forties, fifties, sixties, seventies, and eighties. As the dataset used to train the age recognition model had a small number of samples for the seventies and eighties age groups, resulting in low accuracy for these groups. Speech Emotion recognition models are trained using four popular datasets, named as RAVDESS, SAVEE, TESS and CREMA-D. RAVDESS contains eight emotion classes (Anger, Sad, Fear, Happy, Disgust, Surprise, Calm and Neutral) of speech samples, SAVEE contains Seven emotion classes (Anger, Sad, Fear, Happy, Disgust, Surprise, and Neutral) of speech samples, TESS and CREMA-D contains Six emotion classes (Anger, Sad, Fear, Happy, Disgust, and Neutral) of speech samples, these speech samples are combined together and stored in a new data-frame which will then be used to extract the features. The model to predict emotion using acoustic features, named as Speech Emotion Recognition, is trained using Convolutional Neural Network (CNN) and has achieved an accuracy of 82%. This model is trained to predict eight classes of emotions such as Anger, Sadness, Fear, Happy, Disgust, Surprise, Calm, and Neutral. As the dataset used has more number of speech samples for Anger, Sad, Fear, Happy, Disgust, Neutral than Calm and Surprise makes the model more fitted to those emotions, results in poor prediction of other

emotions. The models predict the emotions anger, calm, fear, sad, and surprise with accuracy more than 85% but disgust, neutral and happy with little low accuracy. On the other hand, the Text Emotion Recognition model is trained by extracting the lexical features from the text samples using the Random Forest algorithm, achieving an accuracy of 89%. This model is trained to predict six classes of emotions such as Anger, Sadness, Joy, Surprise, Love, and Fear. This model predict the emotion by getting the input as text samples and predicts sadness, love and fear with accuracy over 90% but predicts joy, anger and surprise with a little low accuracy .Overall, the Speech Analysis System is an effective tool for predicting various features from speech and text samples. The system uses state-of-the-art machine learning algorithms, such as SVM, CNN, and Random Forest, to achieve high accuracy in predicting age, gender, and emotion. The system has the potential to be used in various applications, such as speech recognition, emotion detection, and language processing. However, the accuracy of the models can still be improved by increasing the size of the dataset and incorporating more advanced features.

CHAPTER 7

CONCLUSION

7.1. CONCLUSION:

The study mentioned that the model trained to predict age group did not perform well compared to other models due to insufficient data. Figure 2 in the study shows that the number of speech samples representing the seventies and eighties age groups is significantly lower than the speech samples for other age groups such as teens to sixties. This insufficient data limits the model's ability to understand the speech data correctly and accurately predict the age groups. On the other hand, the study found that the text emotion recognition model performed better than the speech emotion recognition model. To predict the emotion of a new speech sample, the study proposed converting the speech sample into textual content using speech recognition methods and applying it to the text emotion recognition model. This method could help overcome some of the limitations of using only acoustic features to predict emotions in speech. The proposed speech analysis system successfully predicts the gender and emotion of the speaker, but it only partially predicts the age of the speaker. The limitations in predicting age are due to the insufficient data available for training the model to accurately differentiate between different age groups. Additionally, predicting age from speech can be a challenging task due to the variations in speech patterns among individuals of the same age group. Despite these limitations, the proposed speech analysis system is still a significant advancement in the field of speech analysis. It provides a more comprehensive approach to analyzing speech by considering multiple features simultaneously. The ability to predict gender and emotion from speech has many potential applications in areas such as healthcare, psychology, and marketing. To improve the accuracy of predicting age from speech, future research could focus on collecting more speech samples from individuals in the seventies and eighties age groups. Additionally, developing models that can capture both acoustic and linguistic features of speech could help overcome some of the limitations of using only acoustic features to predict emotions in speech. With further advancements in this field, we can expect more comprehensive and accurate models for analyzing speech data in the future.

7.2. FUTURE WORK:

- In future execution, the model to predict the age of the speaker will be improved to get better accuracy and prediction results by using an dataset which contains partially equal amount of speech samples for every age groups and by implementing different feature extraction process.
- Different data augmentation techniques and feature extraction techniques can be used to train Speech Emotion Recognition model to improve the model to predict eight class of emotions.
- Speech Emotion Recognition model can be trained using different datasets which contains equal amount of speech samples for every emotion class which can help improve the model to predict good results for every emotion class.
- Text Emotion Recognition model can be improved to predict more number of emotion classes using different dataset and by using deep learning architecture. Using deep learning architecture to train model with large text samples can improve prediction results of emotion from text sample
- Models can be trained to predict other characters like accent from the speech samples by extracting the acoustic feature, extracting feature methods can be varied to predict accent from the speech samples.

REFERENCES:

- [1] C. -C. Chen, P. -T. Lu, M. -L. Hsia, J. -Y. Ke and O. T. . -C. Chen, "Gender-to-Age hierarchical recognition for speech," 2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS), Seoul, Korea (South), 2011, pp. 1-4, doi: 10.1109/MWSCAS.2011.6026475.
- [2] E. Pantraki, C. Kotropoulos and A. Lanitis, "Age interval and gender prediction using PARAFAC2 applied to speech utterances," 2016 4th International Conference on Biometrics and Forensics (IWBF), Limassol, Cyprus, 2016, pp. 1-6, doi: 10.1109/IWBF.2016.7449694.
- [3] S. Goyal, V. V. Patage and S. Tiwari, "Gender and Age Group Predictions from Speech Features using Multi-Layer Perceptron Model," 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342434.
- [4] Y. C. Pan, M. X. Xu, L. Q. Liu and P. F. Jia, "Emotion-detecting Based Model Selection for Emotional Speech Recognition," The Proceedings of the Multiconference on "Computational Engineering in Systems Applications", Beijing, China, 2006, pp. 2169-2172, doi: 10.1109/CESA.2006.4281997.
- [5] H. Hu, M. -X. Xu and W. Wu, "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, USA, 2007, pp. IV-413-IV-416, doi: 10.1109/ICASSP.2007.366937.
- [6] C. Jie, "Speech emotion recognition based on convolutional neural network," 2021 International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom, 2021, pp. 106-109, doi: 10.1109/NetCIT54147.2021.00028.
- [7] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017, pp. 2257-2260, doi: 10.1109/WiSPNET.2017.8300161.
- [8] Pervaiz, Mahwish & Khan, Tamim. (2016). Emotion Recognition from Speech using Prosodic and Linguistic Features. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070813.
- [9] M. Sakurai and T. Kosaka, "Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results," 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, 2021, pp. 824-827, doi: 10.1109/GCCE53005.2021.9621810.

- [10] S. Bhosale, R. Chakraborty and S. K. Koppurapu, "Deep Encoded Linguistic and Acoustic Cues for Attention Based End to End Speech Emotion Recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7189-7193, doi: 10.1109/ICASSP40776.2020.9054621.
- [11] S. GB and I. J. Jacob, "A Semantic Approach for Computing Speech Emotion Text Classification Using Machine Learning Algorithms," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichy, India, 2022, pp. 1-5, doi: 10.1109/ICEEICT53079.2022.9768465.
- [12] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto and S. Makino, "Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features," in IEEE Access, vol. 10, pp. 115732-115743, 2022, doi: 10.1109/ACCESS.2022.3219094.
- [13] Shiqing Zhang, "Speech emotion recognition based on Fuzzy Least Squares Support Vector Machines," 2008 7th World Congress on Intelligent Control and Automation, Chongqing, China, 2008, pp. 1299-1302, doi: 10.1109/WCICA.2008.4594449.
- [14] C. Wang, S. Ju, Y. Liu, R. Chen and X. Huang, "Adaptive Feature Extractor of Global Representation and Local Semantics for Text Classification," in IEEE Access, vol. 8, pp. 202687-202695, 2020, doi: 10.1109/ACCESS.2020.3036455.
- [15] R. Lotfian and C. Busso, "Lexical Dependent Emotion Detection Using Synthetic Speech Reference," in IEEE Access, vol. 7, pp. 22071-22085, 2019, doi: 10.1109/ACCESS.2019.2898353.

APPENDIX

A. SOURCE CODE

[1] Gender Recognition Model

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt # visualizing data
import seaborn as sns # visualizing data with stunning default theme
import sklearn # contain algorithms
import warnings
warnings.filterwarnings('ignore')

# load dataset from input directory
df = pd.read_csv('C:/Users/ARAVIND/Downloads/cv-valid-train(csv)/cv-valid-train.csv')
df[df['age'].notna()].head()

sns.set(rc={'figure.figsize':(15, 5)})
sns.countplot(x="age",
              data=df[df['age'].notna()],
              order=['teens', 'twenties', 'thirties', 'fourties', 'fifties', 'sixties', 'seventies', 'eighties'])

plt.show()

#we extract the columns that we think useful are
df = df[['filename', 'age', 'gender']]
#To clean the data we remove the sample with NaN attribute values.
data = df[df['age'].notna() & df['gender'].notna()]
data.reset_index(inplace=True, drop=True)
data.head()

#data['gender'] = pd.to_numeric(data['gender'], errors = 'coerce')
#data.dtypes
#if the below code fail to covert gender datatype and values then uncomment above code
cleanup_nums = {"age":
{"teens":0.0,"twenties":1.0,"thirties":2.0,"fourties":3.0,"fifties":4.0,"sixties":5.0,"seventies":6.0,"
twenties":7.0,"eighties":8.0}}
data = data.replace(cleanup_nums)
data = data[:1000]
data.head()

import librosa
ds_path = "C:/Users/ARAVIND/Downloads/gender data/"

#this function is used to extract audio frequency features
def feature_extraction(filename, sampling_rate=48000):
    path = "{}{}".format(ds_path, filename)
    features = list()
    audio, _ = librosa.load(path, sr=sampling_rate)
```

```

gender = data[data['filename'] == filename].gender.values[0]
spectral_centroid = np.mean(librosa.feature.spectral_centroid(y=audio, sr=sampling_rate))
spectral_bandwidth = np.mean(librosa.feature.spectral_bandwidth(y=audio,
sr=sampling_rate))
spectral_rolloff = np.mean(librosa.feature.spectral_rolloff(y=audio, sr=sampling_rate))
features.append(gender)
features.append(spectral_centroid)
features.append(spectral_bandwidth)
features.append(spectral_rolloff)

mfcc = librosa.feature.mfcc(y=audio, sr=sampling_rate)
for el in mfcc:
    features.append(np.mean(el))

return features

features = feature_extraction(data.iloc[0]['filename'])
print("features: ", features)

#the function create dataframe to store the feature and label related to each other
def create_df_features(orig):
    new_rows = list()
    tot_rows = len(orig)-1
    stop_counter = 55001

    for idx, row in orig.iterrows():
        if idx >= stop_counter: break
        print("\r", end="")
        print("{} / {}".format(idx, tot_rows), end="", flush=True)
        features = feature_extraction(row['filename'])
        features.append(row['age'])
        new_rows.append(features)

    return pd.DataFrame(new_rows, columns=["label", "spectral_centroid",
"spectral_bandwidth", "spectral_rolloff",
        "mfcc1", "mfcc2", "mfcc3", "mfcc4", "mfcc5", "mfcc6", "mfcc7", "mfcc8",
        "mfcc9", "mfcc10", "mfcc11", "mfcc12", "mfcc13", "mfcc14", "mfcc15",
"mfcc16",
        "mfcc17", "mfcc18", "mfcc19", "mfcc20", "age"])

df_features = create_df_features(data)
df_features.head()

from sklearn.preprocessing import StandardScaler

def scale_features(data):
    scaler = StandardScaler()
    scaled_data = scaler.fit_transform(np.array(data.iloc[:, 1:], dtype = float))
    # with data.iloc[:, 0:-1] we don't consider the label column

    return scaled_data, scaler

x, scaler = scale_features(df_features)

print("Before scaling:", df_features.iloc[0].values[:-1])

```

```

print("\nAfter scaling:", x[0])

from sklearn.preprocessing import LabelEncoder

def get_labels(data):
    labels = data.iloc[:, 0]
    encoder = LabelEncoder()
    labels = encoder.fit_transform(labels)
    return labels, encoder

y, encoder = get_labels(df_features)
classes = encoder.classes_
print("Before encoding:", df_features.iloc[0].values[0])
print("\nAfter encoding:", y[0])
print("\nClasses:", classes)

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif

n_features = 22

f_selector = SelectKBest(f_classif, k=n_features).fit(x, y)
X_new = f_selector.transform(x)
scores = f_selector.scores_

indices = np.argsort(scores)[::-1]

features = []
for i in range(n_features):
    features.append(df_features.columns[indices[i]])

plt.figure(figsize=(22, 5))
plt.bar(features, scores[indices[range(n_features)]]], color='g')
plt.xticks(fontsize=8)
plt.show()

from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix

X_train, X_test, y_train, y_test = train_test_split(X_new, y, test_size=0.2, random_state=0)

model = SVC(C=100, gamma='scale')
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

cm = confusion_matrix(y_test, y_pred)

plt.figure()
my_plot_confusion_matrix(cm, classes=classes)

plt.grid(False)
plt.show()

# Plot normalized confusion matrix

```

```

plt.figure()
my_plot_confusion_matrix(cm, classes=classes, normalize=True)

plt.grid(False)
plt.show()

import joblib
# save the model to disk
filename = 'finalized_model.sav'
joblib.dump(model, filename)

from sklearn.metrics import classification_report
cr = classification_report(y_test, y_pred)
print(cr)

```

[2] Speech Emotion Recognition Model

```

import pandas as pd
import numpy as np

import os
import sys

# librosa is a Python library for analyzing audio and music. It can be used to extract the data
# from the audio files we will see it later.
import librosa
import librosa.display
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# to play the audio files
from IPython.display import Audio

import keras
from keras.callbacks import ReduceLROnPlateau
from keras.models import Sequential, load_model
from keras.layers import Dense, Conv1D, MaxPooling1D, Flatten, Dropout,
BatchNormalization, Activation
from keras.utils import np_utils, to_categorical
from keras.callbacks import ModelCheckpoint

import warnings
if not sys.warnoptions:
    warnings.simplefilter("ignore")
warnings.filterwarnings("ignore", category=DeprecationWarning)

# Paths for data.
Ravdess = "C:/Users/hp/Downloads/RAVDESS/audio_speech_actors_01-24/"
Crema = "C:/Users/hp/Downloads/crema-d/AudioWAV/"
Tess = "C:/Users/hp/Downloads/TESS/TESS Toronto emotional speech set data/TESS

```

```

Toronto emotional speech set data/"
Savee = "C:/Users/hp/Downloads/Savee/ALL/"

ravdess_directory_list = os.listdir(Ravdess)

file_emotion = []
file_path = []
for dir in ravdess_directory_list:
    # as there are 20 different actors in our previous directory we need to extract files for each
    # actor.
    actor = os.listdir(Ravdess + dir)
    for file in actor:
        part = file.split('.')[0]
        part = part.split('-')
        # third part in each file represents the emotion associated to that file.
        file_emotion.append(int(part[2]))
        file_path.append(Ravdess + dir + '/' + file)

# dataframe for emotion of files
emotion_df = pd.DataFrame(file_emotion, columns=['Emotions'])

# dataframe for path of files.
path_df = pd.DataFrame(file_path, columns=['Path'])
Ravdess_df = pd.concat([emotion_df, path_df], axis=1)

# changing integers to actual emotions.
Ravdess_df.Emotions.replace({1:'neutral', 2:'calm', 3:'happy', 4:'sad', 5:'angry', 6:'fear',
7:'disgust', 8:'surprise'}, inplace=True)
Ravdess_df.head()

crema_directory_list = os.listdir(Crema)

file_emotion = []
file_path = []

for file in crema_directory_list:
    # storing file paths
    file_path.append(Crema + file)
    # storing file emotions
    part=file.split('_')
    if part[2] == 'SAD':
        file_emotion.append('sad')
    elif part[2] == 'ANG':
        file_emotion.append('angry')
    elif part[2] == 'DIS':
        file_emotion.append('disgust')
    elif part[2] == 'FEA':
        file_emotion.append('fear')
    elif part[2] == 'HAP':
        file_emotion.append('happy')
    elif part[2] == 'NEU':
        file_emotion.append('neutral')
    else:
        file_emotion.append('Unknown')

# dataframe for emotion of files

```

```

emotion_df = pd.DataFrame(file_emotion, columns=['Emotions'])

# dataframe for path of files.
path_df = pd.DataFrame(file_path, columns=['Path'])
Crema_df = pd.concat([emotion_df, path_df], axis=1)
Crema_df.head()

tess_directory_list = os.listdir(Tess)

file_emotion = []
file_path = []

for dir in tess_directory_list:
    directories = os.listdir(Tess + dir)
    for file in directories:
        part = file.split('.')[0]
        part = part.split('_')[2]
        if part=='ps':
            file_emotion.append('surprise')
        else:
            file_emotion.append(part)
        file_path.append(Tess + dir + '/' + file)

# dataframe for emotion of files
emotion_df = pd.DataFrame(file_emotion, columns=['Emotions'])

# dataframe for path of files.
path_df = pd.DataFrame(file_path, columns=['Path'])
Tess_df = pd.concat([emotion_df, path_df], axis=1)
Tess_df.head()

savee_directory_list = os.listdir(Savee)

file_emotion = []
file_path = []

for file in savee_directory_list:
    file_path.append(Savee + file)
    part = file.split('_')[1]
    ele = part[:-6]
    if ele=='a':
        file_emotion.append('angry')
    elif ele=='d':
        file_emotion.append('disgust')
    elif ele=='f':
        file_emotion.append('fear')
    elif ele=='h':
        file_emotion.append('happy')
    elif ele=='n':
        file_emotion.append('neutral')
    elif ele=='sa':
        file_emotion.append('sad')
    else:
        file_emotion.append('surprise')

# dataframe for emotion of files

```

```

emotion_df = pd.DataFrame(file_emotion, columns=['Emotions'])

# dataframe for path of files.
path_df = pd.DataFrame(file_path, columns=['Path'])
Savee_df = pd.concat([emotion_df, path_df], axis=1)
Savee_df.head()

# creating Dataframe using all the 4 dataframes we created so far.
data_path = pd.concat([Ravdess_df, Crema_df, Tess_df, Savee_df], axis = 0)
data_path.to_csv("data_path.csv",index=False)
data_path.head()

def noise(data):
    noise_amp = 0.035*np.random.uniform()*np.amax(data)
    data = data + noise_amp*np.random.normal(size=data.shape[0])
    return data

def stretch(data):
    return librosa.effects.time_stretch(data,rate=0.8)

def shift(data):
    shift_range = int(np.random.uniform(low=-5, high = 5)*1000)
    return np.roll(data, shift_range)

def pitch(data, sampling_rate):
    return librosa.effects.pitch_shift(data , sampling_rate, pitch_factor=0.7)

# taking any example and checking for techniques.
path = np.array(data_path.Path)[1]
data, sample_rate = librosa.load(path)

def extract_features(data):
    # ZCR
    result = np.array([])
    zcr = np.mean(librosa.feature.zero_crossing_rate(y=data).T, axis=0)
    result=np.hstack((result, zcr)) # stacking horizontally

    # Chroma_stft
    stft = np.abs(librosa.stft(data))
    chroma_stft = np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T, axis=0)
    result = np.hstack((result, chroma_stft)) # stacking horizontally

    # MFCC
    mfcc = np.mean(librosa.feature.mfcc(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mfcc)) # stacking horizontally

    # Root Mean Square Value
    rms = np.mean(librosa.feature.rms(y=data).T, axis=0)
    result = np.hstack((result, rms)) # stacking horizontally

    # MelSpectrogram
    mel = np.mean(librosa.feature.melspectrogram(y=data, sr=sample_rate).T, axis=0)
    result = np.hstack((result, mel)) # stacking horizontally

    return result

```

```

def get_features(path):
    # duration and offset are used to take care of the no audio in start and the ending of each
    audio files as seen above.
    data, sample_rate = librosa.load(path, duration=2.5, offset=0.6 , sr = 8025)

    # without augmentation
    res1 = extract_features(data)
    result = np.array(res1)

    # data with noise
    noise_data = noise(data)
    res2 = extract_features(noise_data)
    result = np.vstack((result, res2)) # stacking vertically

    # data with stretching and pitching
    new_data = stretch(data)

    res3 = extract_features(new_data)
    result = np.vstack((result, res3)) # stacking vertically

    return result

X, Y = [], []
for path, emotion in zip(data_path.Path, data_path.Emotions):
    feature = get_features(path)
    for ele in feature:
        X.append(ele)
        # appending emotion 3 times as we have made 3 augmentation techniques on each
        audio file.
        Y.append(emotion)

Features = pd.DataFrame(X)
Features['labels'] = Y
Features.to_csv('features.csv', index=False)
Features.head()

X = Features.iloc[:, :-1].values
Y = Features['labels'].values

# As this is a multiclass classification problem onehotencoding our Y.
encoder = OneHotEncoder()
Y = encoder.fit_transform(np.array(Y).reshape(-1,1)).toarray()

# splitting data
x_train, x_test, y_train, y_test = train_test_split(X, Y, random_state=0, shuffle=True)
x_train.shape, y_train.shape, x_test.shape, y_test.shape

# scaling our data with sklearn's Standard scaler
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
x_train.shape, y_train.shape, x_test.shape, y_test.shape

# making our data compatible to model.
x_train = np.expand_dims(x_train, axis=2)
x_test = np.expand_dims(x_test, axis=2)

```

```
x_train.shape, y_train.shape, x_test.shape, y_test.shape
```

```
model = Sequential()
model.add(Conv1D(256, 8, padding='same', input_shape=(x_train.shape[1],1))) #
X_train.shape[1] = No. of Columns
model.add(Activation('relu'))
model.add(Conv1D(256, 8, padding='same'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(Dropout(0.25))
model.add(MaxPooling1D(pool_size=(8)))
model.add(Conv1D(128, 8, padding='same'))
model.add(Activation('relu'))
model.add(Conv1D(128, 8, padding='same'))
model.add(Activation('relu'))
model.add(Conv1D(128, 8, padding='same'))
model.add(Activation('relu'))
model.add(Conv1D(128, 8, padding='same'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(Dropout(0.25))
model.add(MaxPooling1D(pool_size=(8)))
model.add(Conv1D(64, 8, padding='same'))
model.add(Activation('relu'))
model.add(Conv1D(64, 8, padding='same'))
model.add(Activation('relu'))
model.add(Flatten())
model.add(Dense(y_train.shape[1])) # Target class number
model.add(Activation('softmax'))
model.compile(optimizer = 'adam' , loss = 'categorical_crossentropy' , metrics = ['accuracy'])
model.summary()

rlrp = ReduceLROnPlateau(monitor='loss', factor=0.4, verbose=0, patience=2,
min_lr=0.000001)
history=model.fit(x_train, y_train, batch_size=64, epochs=50, validation_data=(x_test, y_test),
callbacks=[rlrp])

print("Accuracy of our model on test data : " , model.evaluate(x_test,y_test)[1]*100 , "%")
```

[3] Text Emotion Recognition Model

```
import re
import nltk
import string
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer, WordNetLemmatizer
```

```

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split

from tensorflow.keras.utils import to_categorical
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences

from tensorflow.keras.optimizers import Adam
from tensorflow.keras.models import Sequential
from tensorflow.keras.callbacks import EarlyStopping
from tensorflow.keras.layers import Dense, LSTM, Embedding, Bidirectional

nltk.download("stopwords")
stop_words = set(stopwords.words("english"))
lemmatizer= WordNetLemmatizer()

# Modelling
from sklearn.model_selection import train_test_split, KFold, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.pipeline import Pipeline
from sklearn.metrics import f1_score
from sklearn.svm import SVC

#Lime
from lime import lime_text
from lime.lime_text import LimeTextExplainer
from lime.lime_text import IndexedString, IndexedCharacters
from lime.lime_base import LimeBase
from lime.lime_text import explanation
sns.set(font_scale=1.3)
nltk.download('omw-1.4')

# Read datasets
df_train = pd.read_csv('C:/Users/ARAVIND/Downloads/emotionText/train.txt', names=['Text',
'Emotion'], sep=';')
df_val = pd.read_csv('C:/Users/ARAVIND/Downloads/emotionText/val.txt', names=['Text',
'Emotion'], sep=';')
df_test = pd.read_csv('C:/Users/ARAVIND/Downloads/emotionText/test.txt', names=['Text',
'Emotion'], sep=';')

#removing duplicated values
index = df_train[df_train.duplicated() == True].index
df_train.drop(index, axis = 0, inplace = True)
df_train.reset_index(inplace=True, drop = True)

#print the rows which are duplicated (duplicated in the text but with different emotions)
df_train[df_train['Text'].duplicated() == True]

#print some of those rows to check
df_train[df_train['Text'] == df_train.iloc[7623]['Text']]

```

```

df_train[df_train['Text'] == df_train.iloc[14313]['Text']]

df_train[df_train['Text'] == df_train.iloc[13879]['Text']]

#removing duplicated text
index = df_train[df_train['Text'].duplicated() == True].index
df_train.drop(index, axis = 0, inplace = True)
df_train.reset_index(inplace=True, drop = True)

#Count the number of stopwords in the data
temp = df_train.copy()
stop_words = set(stopwords.words("english"))
temp['stop_words'] = temp['Text'].apply(lambda x: len(set(x.split()) & set(stop_words)))
temp.stop_words.value_counts()

#distribution of stopwords visually
temp['stop_words'].plot(kind= 'hist')

#check if the data is balanced or not
df_test.Emotion.value_counts()

def lemmatization(text):
    lemmatizer= WordNetLemmatizer()

    text = text.split()

    text=[lemmatizer.lemmatize(y) for y in text]

    return " ".join(text)

def remove_stop_words(text):

    Text=[i for i in str(text).split() if i not in stop_words]
    return " ".join(Text)

def Removing_numbers(text):
    text=" ".join([i for i in text if not i.isdigit()])
    return text

def lower_case(text):

    text = text.split()

    text=[y.lower() for y in text]

    return " ".join(text)

def Removing_punctuations(text):
    ## Remove punctuations
    text = re.sub('[%s]' % re.escape('!"#$%&()*+,-./:;<=>?@[\\]^_`{|}~""'), ' ', text)
    text = text.replace(';', '')

    ## remove extra whitespace
    text = re.sub('\s+', ' ', text)
    text = " ".join(text.split())

```

```

return text.strip()

def Removing_urls(text):
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    return url_pattern.sub("", text)

def remove_small_sentences(df):
    for i in range(len(df)):
        if len(df.text.iloc[i].split()) < 3:
            df.text.iloc[i] = np.nan

def normalize_text(df):
    df.Text=df.Text.apply(lambda text : lower_case(text))
    df.Text=df.Text.apply(lambda text : remove_stop_words(text))
    df.Text=df.Text.apply(lambda text : Removing_numbers(text))
    df.Text=df.Text.apply(lambda text : Removing_punctuations(text))
    df.Text=df.Text.apply(lambda text : Removing_urls(text))
    df.Text=df.Text.apply(lambda text : lemmatization(text))
    return df

def normalized_sentence(sentence):
    sentence= lower_case(sentence)
    sentence= remove_stop_words(sentence)
    sentence= Removing_numbers(sentence)
    sentence= Removing_punctuations(sentence)
    sentence= Removing_urls(sentence)
    sentence= lemmatization(sentence)
    return sentence

df_train= normalize_text(df_train)
df_test= normalize_text(df_test)
df_val= normalize_text(df_val)

#Preprocess text
X_train = df_train['Text'].values
y_train = df_train['Emotion'].values

X_test = df_test['Text'].values
y_test = df_test['Emotion'].values

X_val = df_val['Text'].values
y_val = df_val['Emotion'].values

def train_model(model, data, targets):
    """
    Train a model on the given data and targets.

    Parameters:
    model (sklearn model): The model to be trained.
    data (list of str): The input data.
    targets (list of str): The targets.

    Returns:
    Pipeline: The trained model as a Pipeline object.
    """

```

```

# Create a Pipeline object with a TfidfVectorizer and the given model
text_clf = Pipeline([('vect',TfidfVectorizer()),
                     ('clf', model)])
# Fit the model on the data and targets
text_clf.fit(data, targets)
return text_clf

def get_F1(trained_model,X,y):
    """
    Get the F1 score for the given model on the given data and targets.

    Parameters:
    trained_model (sklearn model): The trained model.
    X (list of str): The input data.
    y (list of str): The targets.

    Returns:
    array: The F1 score for each class.
    """
    # Make predictions on the input data using the trained model
    predicted=trained_model.predict(X)
    # Calculate the F1 score for the predictions
    f1=f1_score(y,predicted, average=None)
    # Return the F1 score
    return f1

#Train the model with the training data
RF = train_model(RandomForestClassifier(random_state = 0), X_train, y_train)

#test the model with the test data
y_pred=RF.predict(X_test)

#calculate the accuracy
RF_accuracy = accuracy_score(y_test, y_pred)
print('Accuracy: ', RF_accuracy,'\n')

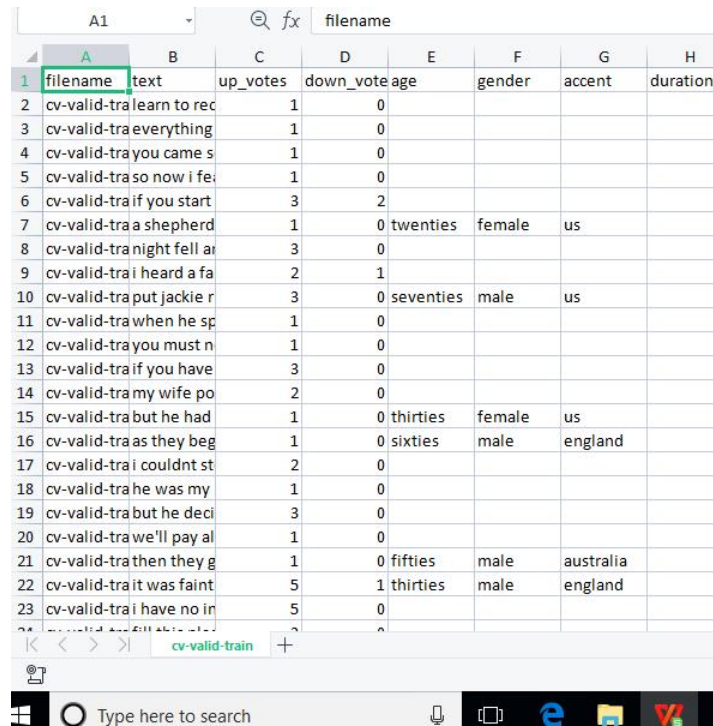
#calculate the F1 score
f1_Score = get_F1(RF, X_test, y_test)
pd.DataFrame(f1_Score, index=df_train.Emotion.unique(), columns=['F1 score'])

##Classification Report
print(classification_report(y_test, y_pred))

```

B. SCREENSHOTS

Snapshots of Gender and Age Recognition models



	A1							
	filename	text	up_votes	down_vote	age	gender	accent	duration
2	cv-valid-tra	learn to rec	1	0				
3	cv-valid-tra	everything	1	0				
4	cv-valid-tra	you came s	1	0				
5	cv-valid-tra	so now i fe	1	0				
6	cv-valid-tra	if you start	3	2				
7	cv-valid-tra	a shepherd	1	0	twenties	female	us	
8	cv-valid-tra	night fell ar	3	0				
9	cv-valid-tra	i heard a fa	2	1				
10	cv-valid-tra	put jackie r	3	0	seventies	male	us	
11	cv-valid-tra	when he sp	1	0				
12	cv-valid-tra	you must n	1	0				
13	cv-valid-tra	if you have	3	0				
14	cv-valid-tra	my wife po	2	0				
15	cv-valid-tra	but he had	1	0	thirties	female	us	
16	cv-valid-tra	as they beg	1	0	sixties	male	england	
17	cv-valid-tra	i couldnt st	2	0				
18	cv-valid-tra	he was my	1	0				
19	cv-valid-tra	but he deci	3	0				
20	cv-valid-tra	we'll pay al	1	0				
21	cv-valid-tra	then they g	1	0	fifties	male	australia	
22	cv-valid-tra	it was faint	5	1	thirties	male	england	
23	cv-valid-tra	i have no in	5	0				

Fig.B.1. dataset used for gender and age model



Fig.B.2. Speech Samples used in Gender and Age model

	label	spectral_centroid	spectral_bandwidth	spectral_rolloff	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	...	mfcc12	mfcc13
0	female	2147.605880	2430.474971	4428.830553	-625.281433	111.306145	6.369085	34.767101	31.623457	-4.721563	...	-10.708405	-11.20651
1	male	2815.325442	2451.922348	4884.633819	-469.936646	126.283386	-16.548649	3.546792	2.184188	13.748076	...	-10.987889	1.13257
2	female	1844.637736	1491.011524	3164.948048	-418.205475	147.666870	-49.974003	-2.286364	37.185310	-8.906048	...	-13.844881	-9.92857
3	male	2123.711333	2202.012929	4111.215965	-464.910706	118.437225	19.749662	27.143229	26.438824	2.309232	...	-5.461425	-2.90847
4	male	2360.672044	2957.220240	4767.080050	-343.833008	157.153885	7.661150	41.898964	-15.152472	28.925106	...	0.102647	6.31981

5 rows x 25 columns

Fig.B.3. Features extracted from the speech sample for Gender

	precision	recall	f1-score	support
0	0.74	0.67	0.70	46
1	0.90	0.93	0.91	153
2	0.00	0.00	0.00	1
accuracy			0.86	200
macro avg	0.55	0.53	0.54	200
weighted avg	0.86	0.86	0.86	200

Fig.B.4. Classification report of Gender Model

	precision	recall	f1-score	support
0	0.27	0.18	0.22	22
1	0.49	0.63	0.55	70
2	0.31	0.39	0.34	38
3	0.60	0.43	0.50	28
4	0.50	0.38	0.43	21
5	0.78	0.50	0.61	14
6	1.00	0.29	0.44	7
accuracy			0.46	200
macro avg	0.56	0.40	0.44	200
weighted avg	0.49	0.46	0.46	200

Fig.B.5. Classification report of Age Model

Screenshots of Speech Emotion Recognition model

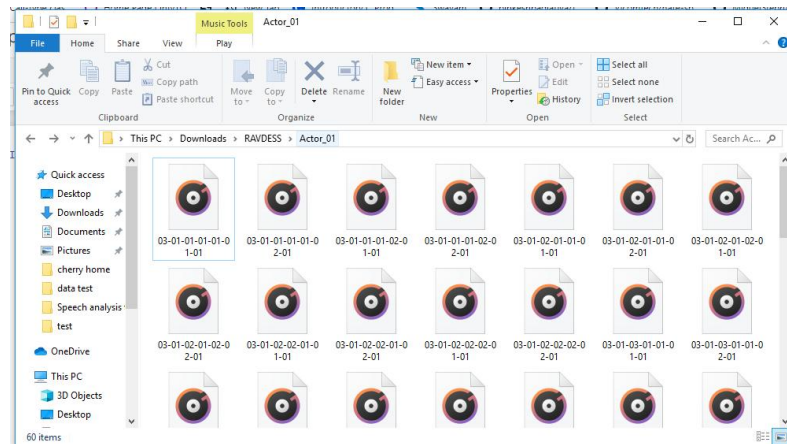


Fig.B.6. Speech sampled dataset of Speech Emotion Model

	0	1	2	3	4	5	6	7	8	9	...	153	154	155	156	157
0	0.303162	0.708741	0.733662	0.784873	0.768437	0.673473	0.650822	0.704595	0.634349	0.612953	...	0.000091	0.000128	0.000172	0.000145	0.000211
1	0.318372	0.701556	0.739939	0.810096	0.829600	0.790828	0.725413	0.727147	0.732578	0.649151	...	0.000109	0.000146	0.000192	0.000166	0.000229
2	0.171626	0.660387	0.687041	0.771159	0.806366	0.774695	0.610904	0.674279	0.712011	0.628345	...	0.000035	0.000056	0.000088	0.000070	0.000098
3	0.330127	0.638073	0.676395	0.750561	0.729686	0.692049	0.624928	0.668495	0.673284	0.576035	...	0.000076	0.000122	0.000094	0.000137	0.000184
4	0.306909	0.687581	0.737560	0.783938	0.768706	0.724998	0.669270	0.702490	0.671643	0.607457	...	0.000088	0.000129	0.000107	0.000153	0.000199

5 rows x 163 columns

Fig.B.7. Extracted features for Speech Emotion Model

	precision	recall	f1-score	support
angry	0.92	0.90	0.91	1438
calm	0.89	0.63	0.74	137
disgust	0.68	0.89	0.77	1468
fear	0.86	0.78	0.82	1424
happy	0.78	0.89	0.83	1462
neutral	0.83	0.76	0.80	1310
sad	0.90	0.68	0.78	1400
surprise	0.90	0.94	0.92	483
accuracy			0.82	9122
macro avg	0.85	0.81	0.82	9122
weighted avg	0.84	0.82	0.82	9122

Fig.B.8. Classification report of Speech Emotion Model

	Predicted Labels	Actual Labels
0	disgust	disgust
1	disgust	disgust
2	angry	angry
3	disgust	disgust
4	fear	fear
5	fear	fear
6	happy	happy
7	happy	happy
8	sad	sad
9	sad	sad

Fig.B.9. Prediction results of Speech Emotion Model

Screenshots of Text Emotion Recognition model

	Text	Emotion
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so damned...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the fireplac...	love
4	i am feeling grouchy	anger

Fig.B.10. Labeled Dataset of Text Emotion Model

	precision	recall	f1-score	support
anger	0.90	0.91	0.90	275
fear	0.85	0.88	0.87	224
joy	0.89	0.93	0.91	695
love	0.80	0.69	0.74	159
sadness	0.95	0.92	0.93	581
surprise	0.68	0.59	0.63	66
accuracy			0.89	2000
macro avg	0.85	0.82	0.83	2000
weighted avg	0.89	0.89	0.89	2000

Fig.B.11. Classification Report of Text Emotion Model

Screenshot of Prediction Results of New input with all Three Model

```
In [9]: gender_result = gender(data)
age_result = age(data)
text_result = text(data)
speech_result = speech(data)
```

```
Male
FIFTIES
['sadness']
sad
```

Fig.B.12. Prediction Result of New Input

Speech Analysis System using Speech Recognition and Emotion Recognition Methods

Evangelin E
UG Scholar
Department of IT
Sathyabama Institute of Science
and Technology
Chennai, India
evangelinjully@gmail.com

Dr. R Suhashini
Professor
Department of IT
Sathyabama Institute of Science
and Technology
Chennai, India
subhashinigopi@gmail.com

Abstract

As automation's are increasing in this developing world, the need for a machine to understand human speech and their emotions has become an important feature , and also in the field of customer services providers like call centers are started using speech analysis methods to fetch useful in-formations from the speech recordings such as the demographic characteristic and the emotional state of the speaker and the verbal content from the speech. This paper proposes a speech analysis system to analyse the input speech using the speech recognition methods and to predict the gender, age and emotional state of the speaker. Four models are trained : (i) Gender Prediction model based on acoustic features to predict gender as Male, Female or Other, (ii) Age Prediction model based on acoustic features to predict age group as Teens, Twenties, Thirties, Forties, Fifties, Sixties and Seventies, (iii) Speech Emotion Recognition model based on acoustic features to predict Eight class of emotions such as Anger, Sad, Fear, Happy, Surprise, Neutral, Calm and Disgust and (iv) Text Emotion Recognition model based on lexical feature to predict Seven class of emotions such as Anger, Sad, Fear, Joy, Surprise, Neutral and Disgust.

Keywords - *Speech Emotion Recognition , Speech Analysis System , Text emotion recognition , Machine Learning , Gender and age prediction.*

I. INTRODUCTION

Speech analysis is the process of analyzing voice recordings using speech recognition softwares to fetch valuable in-formations. Speech analysis tools analyse the words and audio patterns from the recordings and identifies the demographic characteristics, emotional state and the

content of speech expressed by the speaker. Speech analysis tools are potentially applied in fields like Education, Healthcare, Law Enforcement, Market research and many more fields which has a need to analyse the voice recording and interactions of humans through speech, this makes speech analysis as an important topic of research in recent years. As speech is the main mode of communication for humans, many developing automated technologies started including voice activated assistance as a main feature, which again makes speech analysis as an hot research topic. The speech analysis system proposed in this paper analyse the voice recordings and predict the demographic characters such as age and gender of the speaker and the emotion expressed by the speaker. The age and gender predicting models is trained to make predictions by extracting acoustic features from the speech signal. To predict the emotional state of the speaker, two models are developed : Speech Emotion Recognition model uses the acoustic features of the speech signals to train and predict the emotion , Where as the another model Text Emotion Recognition uses the lexical features of the speech, to train and predict the emotion. The models to predict gender, age and emotion of the speaker using acoustic feature are trained using large data-set of speech samples and developed based on speech recognition methods, whereas the model using lexical feature to predict emotion developed using Natural Language Processing methods and uses a text data-set which contain large number of sentences. The data-sets used to develop the proposed speech analysis system are labeled data-sets and the models are trained using supervised learning approach. The models are trained with both machine learning and deep learning algorithms and Mel Frequency Cepstral Coefficient(MFCC) is used to extract the acoustic features from the speech samples.

II. LITERATURE SURVEY

Many researches have been carried out in the field of speech analysis and speech recognition, to find the demographic characteristics like age and gender from the speech using their acoustic features and emotion state from

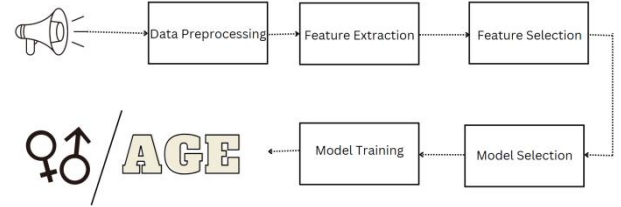
the speech using both acoustic and lexical features. Some of the insights from those researches are taken to improve our project.

Many strategies to predict gender and age from speech have been studied from [1] to [3]. In [1] the author uses an hierarchical approach to predict the age and gender, this approach first predicts the speakers age and uses the gender information to predict age group as male child ,male adult ,male elder and female child, female adult, female elder respectively. As the feature extraction part, he explains about seven speech physical features such as fundamental frequency, Zero Crossing Rate, Duration, Spectral Centroid, Spectral Spread, FFT coefficient and Jitter. The model developed using the proposed hierarchical approach in [1], has low accuracy and prediction results. The system proposed in this paper[2] uses an unified framework to identify the age and gender using Parallel Factor Analysis (PARAFAC2), gives good results for gender prediction but fails for age prediction. In [3], the gender and age prediction models are trained using MLP and the features are extracted using MFCC, formants, pitch, chroma etc and are selected using PCA and Redundant Feature Elimination Technique for preventing the model from over fitting.

In [4], the author tries to improve the speech recognition model considering the emotional hints in the speech, he trained an speech emotion recognition model to recognize emotion by extracting the speech features using MFCC and trained the model using GMM to recognize five classes of emotion, which then used to select speech recognition model. Also in [5], the proposed speech emotion recognition system is trained by considering the spectral features of the speech and uses standard GMM and GMM super-vector based SVM to predict Five class of emotions. The author suggests that gender information should be considered to train emotion recognition model and GMM super-vector based SVM performs better than standard GMM. This paper [6] contains an detailed explanation of feature extraction process using MFCC and how the Convolutional Neural Network architecture can be used to train emotion recognition model in order to get good prediction results. Feature extraction process using MFCC is the main motive of [7], explains how MFCC works and the steps involved in the process of extracting features as Frame Blocking, Windowing, Fast Fourier Transform, Mel Frequency Warping and Mel Frequency Cepstrum Coefficient and using mean and standard deviation for decision making. Emotions are recognized using prosodic, temporal and lexical features from speech[8], two staged model is proposed, to extract emotions from prosodic considering Pitch, Zero Crossing Rate and from temporal features considering MFCC and LPC as first stage and uses lexical feature by implementing an semantic approach to extract emotions from the words spoken as second stage. In [9], the author performs emotion recognition on Japanese language from both acoustic and lexical feature. The work of this author contains information more about recognizing emotion using lexical feature by considering the emotional features as eVector.

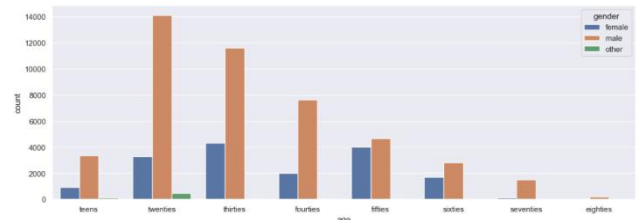
III. GENDER AND AGE RECOGNITION

Recognizing the demographic characters like gender and age of a person from their face features become an easy task as many researches has taken out in that field and gradually improved a lot, but recognizing from their speech signal is an challenging task, here we have mentioned a method by extracting the acoustic feature from the speech signals to predict the age and gender of the speaker. Some of the important steps involved are explained briefly,



A. DATASET:

Very few datasets are openly available with speech samples and their related gender and age in-formations. In this proposed system we have used Mozilla Common Voice data-set, which contains large number of speech samples, but only a small number of speech samples contains their related demographic features like gender, age, accent and . We managed to extract only those samples that contains the aforementioned demographic features and applied in our model to predict age and gender. The below mentioned chart explains the available samples of age groups with their gender information.



B. FEATURE EXTRACTION:

Feature extraction is the process of extracting the features from anything like images or audio signals which then will be used to train the machine learning model to predict the desired results. Here we extract the acoustic characteristics from the speech signals to train our ML model to predict gender and age of the speaker. We extract four types of features from the speech signal with 22 parameters to train our model. The types features extracted from speech signals are explained below,

1. Spectral Centroid: It indicated the center of mass of the frequencies distributed in the audio signal.

2. Spectral Bandwidth: It is derived after the spectral centroid that is the range of frequency in the audio signal around the centroid.

3. Spectral Rolloff: It is the point where a specified percentage of total spectral energy lies.

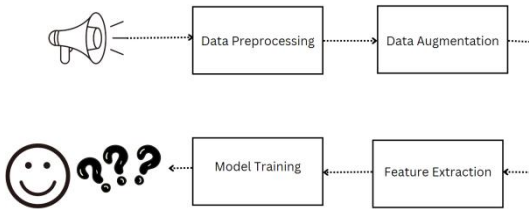
4. MFCC: Mel Frequency Cepstral Coefficient(MFCC) are a set features that are extracted and used most commonly in the field of speech recognition, these features from MFCC are derived from the Mel Frequency cepstrum, by taking logarithm of the power spectrum and transforming to the Mel frequency scale and then applying DCT to obtain cepstral coefficient.

C. MODEL TRAINING

The extracted features using aforementioned methods along with the gender information are used to train the model to predict the age of the speaker, same as the model to predict the gender, uses the extracted features and the age group information to train the model. As explained above, the models to predict age and gender are developed separately using the same extracted feature but as an extra feature the age group is considered for gender recognition model and gender information is considered for age group recognition. The performance of Support Vector Machine and Random Forest on the extracted features are evaluated. SVM performs well than Random Forest, so SVM is used to train the models to predict age and gender.

IV. SPEECH EMOTION RECOGNITION

Recognizing the emotion from the speech signal i.e., using the acoustic characteristics is an complicated task, here in this paper we propose a method to recognize the emotion by including data augmentation process and we tried to extract Eight class of emotions such as Anger, Sad, Fear, Happy, Disgust, Surprise, Calm and Neutral. The below mentioned figure explains the process of Speech Emotion Recognition model.



A. DATASET

Many labeled datasets are available openly with speech samples and their related emotional details expressed in the speech. The proposed Speech Emotion Recognition model is trained using four popular datasets, named as RAVDESS, SAVEE, TESS and CREMA-D. RAVDESS contains eight emotion class (Anger, Sad, Fear, Happy, Disgust, Surprise, Calm and Neutral) of speech samples, SAVEE contains Seven emotion class (Anger, Sad, Fear, Happy, Disgust, Surprise, and Neutral) of speech samples, TESS and CREMA-D contains Six emotion class (Anger, Sad, Fear, Happy, Disgust, and Neutral) of speech samples, these speech samples are combined together and stored in a new data-frame which will then be used to extract the features. The below figure shows the available speech samples of different emotions from the combined data-frame.



B. DATA AUGMENTATION

Data Augmentation is commonly used in machine learning to improve the training dataset in order to prevent the model from over-fitting. In audio processing, data augmentation is applied to audio samples using various transformation techniques to create new samples which then be used to train the model. Here in this proposed model, we used four augmentation techniques to transform the training data, those techniques are mentioned below,

- 1. Noise:** it simply adds random noise to audio signal by calculating its maximum amplitude value then multiplying it with some random value, gives the noise amplitude which then be added to the audio signal.
- 2. Stretch:** It is used to stretch or compress the audio signal, here we compress the audio signal to 20%.
- 3. Shift:** transforms the audio signal by shifting certain number of samples to left or right.
- 4. Pitch:** The pitch of the audio signal is shifted using certain factor, here we shift down the pitch by 30%.

C. FEATURE EXTRACTION

As mentioned above in the gender recognition model, feature extraction is the process extracting the acoustic features from the speech signals. Here in this Speech Emotion Recognition model those features are used to train the DL model that predicts the emotion of the speaker. As we are using Deep Learning architecture, after augmenting the training dataset, five types of features are extracted from the speech signal with 161 parameters to train our model. The types of features extracted for Speech Emotion Recognition model are explained below,

- 1. Zero Crossing Rate:** ZCR is used to estimate the temporal properties of audio signal such as their rhythm or tempo and also this feature is used to differentiate the voiced and unvoiced segments of speech.
- 2. Chroma STFT:** This feature is calculated using the Short Time Fourier Transform(STFT) of the audio signal and used to identify tonal characteristics of speech signal.
- 3. Mel Frequency Cepstral Coefficient:** As mentioned in the gender recognition model, MFCC is the most commonly used feature extraction technique in speech recognition and used to capture the timbre characteristics of speech signal.
- 4. Root Mean Square:** It is used to measure the total energy present in the audio signal by taking square roots of the mean of the squared values of the audio sample. It is

also an most commonly used method, which provides the information about the loudness of the audio signal.

5. Mel Spectrogram: It is used to capture the distribution of energy in the audio signal across different frequency bands, which then can be used to train the ML model.

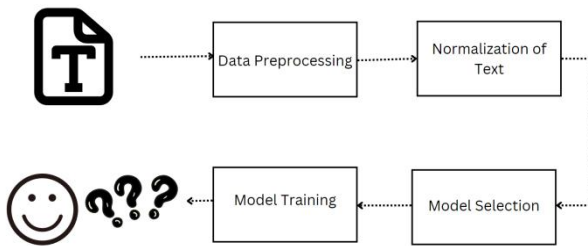
The above mentioned feature extraction methods are used to extract the feature from the speech samples.

D. MODEL TRAINING

After extracting the feature from the speech sample, we used those features to train our Deep Learning Model. As large number of features are extracted from the speech samples, CNN architecture is used to train our model. Usually Convolutional Neural Network requires large dataset to train and are often used in Image processing but as mentioned in [6], CNN can also be effective in Audio Processing. So we used an CNN architecture to train our model that can predict Eight class of emotions such as Anger, Sad, Fear, Happy, Disgust, Surprise, Calm and Neutral.

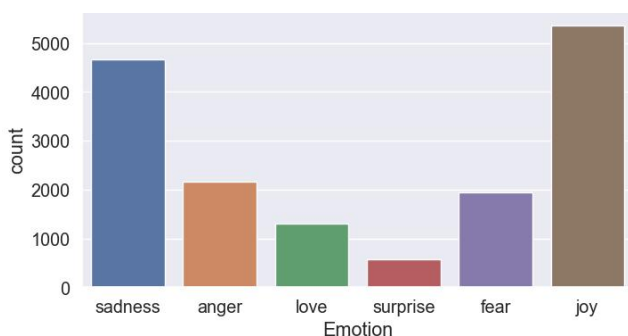
V. TEXT EMOTION RECOGNITION

This model is trained to predict the emotion of the speaker by considering the lexical features i.e, the words or content expressed by the speaker. This model is trained using Natural Language Processing Techniques with text dataset to predict Six class of emotions such as Anger, Sad, Joy, Disgust, Surprise and Fear.



A. DATASET

We have used an open sourced NLP dataset that contains Large number of text sentences and their related emotion as their label. This dataset is separately available for Training, Testing and Validation parts and are specially built and mentioned be used in an model to classify the emotion using the textual content. The below mentioned chart shows the available data in training dataset.



B. NORMALIZING TEXT

In this process we tried to remove the unwanted informations from the data, which will be used for training and testing the model. We use functions to lemmatize words, converts all letters to smaller case and to remove numbers, punctuation, URLs, small sentences and stopwords, stopwords are some commonly used words like 'a', 'an', 'the' 'I' etc. , are removed using a list of predefined stopwords. Applying all these function to remove and convert the words makes the texts more easier for the machine to understand and learn to train the model. This process is mentioned as normalization of the text. We apply this to our dataset before training the model and apply this when we test the model with new data.

C. MODEL TRAINING

After normalizing the text data, we apply that to train our model. We have tried to evaluate the performance of four model that are trained using Machine Learning algorithms like Logistic Regression, Decision tree, Support Vector Machine and Random Forest, to find the best model with better results. Among the four trained model, Random Forest has the best results and are used in the final model to predict the six class of emotions such as Anger, Sad, Fear, Disgust, Surprise and Joy from the textual content expressed by the speaker in the speech.

VI. RESULTS

The above mentioned models trained using the proposed methods are used in our speech analysis system to predict the gender, age and emotion of the speaker. The model trained to predict the gender of the speaker using the proposed method achieved an accuracy above 80%, but the model trained to predict the age group attained an accuracy of 46%, but the model is trained to predict the age groups as Teen, Twenties, Thirties, Forties, Fifties, Sixties, Seventies and Eights. And the Speech emotion recognition model trained using CNN achieved an accuracy about 87%, where as the Text Emotion Recognition model trained using Random Forest achieved 89%, which the best among all trained models.

VII. CONCLUSION

As mentioned in the results the model trained to predict age group got poor results comparing to other model due to insufficient data, as mentioned in the fig.2 the speech samples representing the seventies and eighties age groups are comparatively very less than the speech samples for teens to sixties. This stops the model from understanding the data properly to predict the age groups. Text Emotion Recognition model had better results than Speech Emotion Recognition model, to predict the new speech sample using Text Emotion Recognition model, we convert the speech sample into textual content using Speech Recognition methods and apply them into the model to predict the emotion. Hence, the proposed speech analysis system can Succesfully predict the gender and emotion of the speaker, but partially predicts the age of the speaker.

VIII. REFERENCES

- [1] Chih-Chang Chen, Ping-Tsung Lu, Meng-Lin Hsia, Jia-You Ke and Oscar T-C. Chen ,”Gender-to-Age Hierarchical recognition for speech” ,IEEE 2011.
- [2] Evangelia Pantraki, Constantine Kotropoulus and Andreas Lanitis , “Age interval and Gender Prediction using Parafac2 applied to Speech Utterances” ,IEEE 2016.
- [3] Saksham Goyal,Vinay Vasanth Patage and Sourabh Tiwari, “Gender and Age Group Predictions from Speech Features using Multi-Layer Perceptron Model”, IEEE 2020.
- [4] Y. C. Pan , M. X. Xu, L. Q. Liul, and P. F. Jia, ”Emotion-detecting Based Model Selection for Emotional Speech Recognition” ,IEEE 2006.
- [5] Hao Hu, Ming-Xing Xu, and Wei Wu, ”GMM Supervector based SVM with spectral features for Speech Emotion Recognition”, IEEE 2007.
- [6] Chen Jie ,“Speech Emotion Recognition based on Convolutional Neural Network” , IEEE 2021.
- [7] M.S. Likitha, Sri Raksha R. Gupta, K. Hasitha and A. Upendra Raju ,” Speech Based Human Emotion Recognition Using MFCC” , IEEE 2017.
- [8] Mahwish Pervaiz and Tamim Ahmed Khan ,“Emotion Recognition from Speech using Prosodic and Linguistic Features”, IEEE 2016.
- [9] Misaki Sakurai and Tetsuo Kosaka, “Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results ”, IEEE 2021