

# **SPEECH EMOTION RECOGNITION OF MACHINE LEARNING TECHNIQUES USING PYTHON**

**Naima Farah Mohamed**

**Maryama abdi nasir dahir**

**Shukri Hassan deck**

**Abdirisaq Ise Hussein**

**SUBMISSION OF GRADUATION PROJECT FOR  
PARTIAL FULFILLMENT OF THE  
DEGREE OF BACHELOR OF SCIENCE IN  
COMPUTER APPLICATIONS**

**JAMHURIYA UNIVERSITY OF SCIENCE AND  
TECHNOLOGY (JUST)**

**FACULTY OF COMPUTER & INFORMATION  
TECHNOLOGY**

**AUGUST 2023**

**JAMHRURIYA UNIVERSITY OF SCIENCE AN TECHNOLOGY (JUST)  
ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate 1: Maryam Abdi Nasir Dahir ID No: C119441

Name of Candidate 2: Naima Farah Mohamed ID No: C119443

Name of Candidate 3: Abdirizak Ise Hussein ID No: C119255

Name of Candidate 3: Shukri Hassan Madowe ID No: C119450

Name of Degree: **BACHELOR OF COMPUTER APPLICATION**

Title of Project Paper/Research Report/Dissertation/Thesis (“this Work”):

**SPEECH EMOTION RECOGNITION OF MACHINE LEARNING  
TECHNIQUES USING PYTHON**

Field of Study: Machine Learning

We do solemnly and sincerely declare that:

- 1) We are the sole author of this work;
- 2) This work is original;
- 3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- 4) We do not have any actual knowledge nor do we ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- 5) We hereby assign all and every rights in the copyright to this Work to Jamhuriya University Of Science and technology (“JUST”), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of JUST having been first had and obtained;
- 6) We are fully aware that if in the course of making this Work we have infringed any copyright whether intentionally or otherwise, we may be subject to legal action or any other action as may be determined by JUST.

Candidate 1’s Signature: \_\_\_\_\_ Candidate 2’s Signature: \_\_\_\_\_

Candidate 3’s Signature: \_\_\_\_\_ Candidate 4’s Signature: \_\_\_\_\_

Subscribed and solemnly declared before,

Name of the Supervisor:

Supervisor’s signature: \_\_\_\_\_ Date: \_\_\_\_\_

Witness’s Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Name: \_\_\_\_\_ Designation: \_\_\_\_\_

## **Dedication**

I would like to dedicate this thesis to my parents, who gave the little they had to ensure I would have the opportunity of an education. Their efforts and struggles have allowed me to have a key to unlock the mysteries of our world, and beyond.

## **Abstract**

Over the last few decades, there has been a tremendous amount of research on the use of machine learning for speech processing applications, particularly speech recognition. However, in recent years, research has concentrated on using deep learning for speech-related applications. This new machine area Learning has produced far superior results in a variety of applications, including speech, when compared to others, and has thus become a very appealing area of research. For this we use neural networks methods, our proposed classifier makes easy for the speaker to get their emotions, our classifier predicts the speaker emotions entered manually. The result showed algorithms is high Accuracy; an artificial neural network learning algorithm successfully correctly classified instances as 98.73% percent.

## **Acknowledgement**

First and foremost, we are grateful for the prayers of our mothers, fathers, and our classmates unwavering love and unwavering support throughout our lives as well as our studies. Without you, this thesis project would not have been accomplished, and you have truly molded us into the people we are today. We would like to thank the Almighty ALLAH for the gift of life, intelligence, and understanding that he has bestowed upon us, as well as the cause for our being. And to our families for the love and support they had provided throughout our life. Also, we would like to express our sincere appreciation and gratitude to our research supervisor Bashir Abdirahman whom we regard as our mentor and supervisor, we thank him for the expertise and intelligence he has displayed while supervising this project. This excellent work, we believe, is a result of his excellent supervision and cooperation. Finally, we'd want to thank our Faculty's instructors on their outstanding work during the course of our four-year curriculum. May the Lord bless and keep them safe.

## TABLE OF CONTENTS

<b>Dedication .....</b>	<b>iii</b>
<b>Abstract.....</b>	<b>iv</b>
<b>Acknowledgement .....</b>	<b>v</b>
<b>Table of Contents .....</b>	<b>vi</b>
<b>List of Figures.....</b>	<b>ix</b>
<b>List of Tables .....</b>	<b>x</b>
<b>Chapter I: Introduction.....</b>	<b>1</b>
1.0 Introduction .....	1
1.1 Background of the study .....	1
1.2 Problem statement.....	3
1.3 objective study .....	3
1.4 research questions .....	4
1.5 Scope of the study .....	4
1.6 Significance of the project .....	4
1.7 Organization of the Study .....	5
<b>Chapter II: Literature work .....</b>	<b>7</b>
2.0 introduction .....	7
2.1 History of machine learning.....	7
2.3 Types of machine learning .....	8
2.3.1 Supervised learning.....	8
2.3.2 Unsupervised learning.....	9
2.3.3 Reinforcement learning.....	10
2.4 What is voice recognition (speaker recognition)?.....	10
2.4.1 How voice recognition works .....	11

2.5 Sensory modalities for emotion expression .....	14
2.5.1 Facial expressions .....	14
2.5.2 Speech .....	14
2.5.3 Physiological signals .....	14
2.6 Related Work .....	15
2.7 conclusion .....	26
<b>Chapter III: RESEARCH METHODOLOGY.....</b>	<b>27</b>
3.0 Introduction .....	27
3.1 System Description .....	27
3.2 System architecture .....	27
3.3 System features .....	28
3.4 System requirement.....	30
3.4.1 Hardware Component .....	30
3.4.2. Software component .....	30
<b>CHAPTER IV SYSTEM ANALYSIS AND DESIGNS .....</b>	<b>33</b>
4.1 Introduction .....	33
4.2 System Analysis .....	33
4.3 System current.....	33
4.3.1 Current drawbacks .....	33
4.4 The Need of This Work.....	34
4.5 System proposed .....	34
4.6 System requirements .....	34
4.6.1 Functional Requirements .....	35
4.5.2 Non-Functional Requirements .....	35
4.6 System design .....	36
4.6System Dataset.....	37
<b>Chapter V: Implementation and Testing .....</b>	<b>39</b>

5.1 introduction .....	39
5.2 Overview of the implementation environment .....	39
5.3 Data Collection & Understanding Process.....	40
5.4 Data pre-processing.....	41
5.5 Training and Test Data.....	41
5.6 Classification Algorithms Accuracy Rate.....	43
5.6 Model Deployment.....	43
5.7 System Snapshots.....	43
5.7.1 Login page.....	44
5.7.2 Dashboard page.....	45
<b>Chapter VI: Discussion .....</b>	<b>47</b>
6.1 Introduction .....	47
6.2 Results .....	47
6.3 limitations.....	48
<b>Chapter VII: Conclusion and Future Work .....</b>	<b>50</b>
7.1 introduction .....	50
7.2 Conclusion .....	50
7.3 Recommendation for Future work .....	50
<b>References .....</b>	<b>52</b>



## LIST OF FIGURES

FIGURES	PAGE
1Fig 2.1: List of Development in Technologies used for Speech Recognition .....	17
2Figure 2.2: ML approach for emotion recognition.....	22
3Figure 2.3: Process of Recognizing Emotion from Speech .....	25
4Figure 3.1 System Architecture .....	28
5Figure 4.1 follow chart diagram.....	37
6Figure 5.1 libraries .....	40
7Figure 5.2 Data collection .....	40
8Figure 5.3 Model Summary .....	41
9Figure 5.4 train accuracy and Val accuracy graph.....	42
10Figure 5.5 train loss and Val loss .....	42
11Figure 5.6 Login Page .....	44
12 Figure 5.7choose file/ upload audio .....	45
13 Figure 5.8 result page .....	46

## List of Tables

Table 2.1 previous work .....	18
Table 5.1 Accuracy Rate.....	44

## **CHAPTER I: INTRODUCTION**

### **1.0 Introduction**

Speech is one of the primary means of communication among human beings. Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others.

### **1.1 Background of the study**

Over the last few decades, there has been a tremendous amount of research on the use of machine learning for speech processing applications, particularly speech recognition. However, in recent years, research has concentrated on using deep learning for speech-related applications. This new machine area Learning has produced far superior results in a variety of applications, including speech, when compared to others, and has thus become a very appealing area of research.(Nassif et al., 2019)

In earlier days, people used speech as a means of communication or the way a listener is conveyed by voice or expression. But the idea of machine learning and various methods are necessary for the recognition of speech in the matter of interaction with machines. With a voice as a bio-metric through use and significance, speech has become an important part of speech development. In this article, we attempted to explain a variety of speech and emotion recognition techniques and comparisons between several methods based on existing algorithms and mostly speech-based methods(Professor, Sreyas Institute of Engineering and Technology, Hyderabad, India. et al., 2020)

Recognizing emotions from voice signals is known as speech emotion recognition.

In order to advance human-computer interaction, the following is crucial:

The five main research areas in human computer interaction are as follows: Research on matching models, task-level research, design, and organizational influence, as well as research on task-level interface hardware and software.

Automatic SER aids virtual assistants and smart speakers with better understanding its users, particularly when they detect phrases with ambiguous meaning. For instance, the word "truly" can be used both positively and negatively to underline and stress out a statement or to cast doubt on a reality. Read the following phrases several times: I like having that gadget a lot. The same program can translate between languages, which is especially useful given that different languages have distinct conventions for expressing emotions in speech (Abbaschian et al., 2021)

Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Research has revealed the powerful role that emotion plays in shaping human social interaction. Emotional displays convey considerable information about the mental state of an individual. This has opened up a new research field called automatic emotion recognition, having basic goals to understand and retrieve desired emotions. In prior studies, several modalities have been explored to recognize the emotional states such as facial expressions, speech, physiological signals.(Kerkeni et al., 2020).

Speech analysis has become a crucial tool in closing the gap between the real and digital worlds as human-machine interaction increases.

The recognition of emotion in speech signals, which was previously researched in linguistics and psychology, is a significant subfield within this subject.

There are several uses for speech emotion recognition. The main goal of this essay is to identify spoken emotions and categorize them into seven different emotion output classes:

anger, boredom, disgust, anxiety, happiness, sorrow, and neutral. Research studies have provided evidence that human emotions in fluency the decision making process to a certain e extent (Ghai et al., 2017)

## **1.2 Problem statement**

It is known that some physiological changes occur in the body due to people's emotional state. Some variables such as pulse, blood pressure, facial expressions, body movements, brain waves, and acoustic properties vary Often in the interest to increase the acceptability of speech technology for human users. The speech signal communicates linguistic information between speakers as well as paralinguistic information about speaker's emotions personality's attitude feelings levels of stress and current mental states.

depending on the emotional state. Pulse, blood pressure, brain waves, and so forth. Although changes cannot be detected without a portable medical device, facial expressions and voice signals can be received directly without connecting any device to the person.

Words are not enough to correctly understand the mood and intention of speaker and thus the introduction if human social skills to human machine communication is of paramount importance. This can be achieved by the researching and creating methods of speech modeling and analysis that embrace to signal linguistic and emotional aspects of communication. In this study we are going to develop a speech recognition system that can detect the emotions conveyed in human speech.

## **1.3 objective study**

The main objectives of this study are:

- To collect and prepare a dataset of speech recordings labelled with different emotions.

- To train and optimize the chosen machine learning algorithm(s) on the dataset, achieving high accuracy in emotion classification.
- To develop a real-time speech emotion recognition system that can be integrated into existing human-computer interaction systems.
- To explore the potential applications of the developed system, such as in call centers, customer service centers, and human-robot interaction systems.

#### **1.4 research questions**

- How to train and optimize the chosen machine learning algorithm(s) on the dataset, achieving high accuracy in emotion classification?
- How to develop a real-time speech emotion recognition system that can be integrated into existing human-computer interaction systems?

#### **1.5 Scope of the study**

Emotions recognition has wide scope in many areas such as human computer interaction, biometric security Etc.

This study will define the purpose of the study speech recognition through emotions. So it provides insight into artificial intelligence or machine intelligence that uses various supervised and unsupervised machine learning algorithms to simulate the human brain

This project's geographical scope or context is in Mogadishu, Somalia. This study will take for nearly a year from October 2022 to July 2023.

#### **1.6 Significance of the project**

The following systems can be cited as an example of the areas in which these studies are used and their intended use:

Education: a course system for distance education can detect bored users so that they can change the style or level of material provided in addition, provide emotional incentives or compromises.

Automobile: driving performance and the emotional state of the driver are often linked internally. Therefore, these systems can be used to promote the driving experience and to improve driving performance.

Security: They can be used as support systems in public spaces by detecting extreme feelings such as fear and anxiety.

Communication: in call centers, when the automatic emotion recognition system is integrated with the interactive voice response system, it can help improve customer service.

Health: It can be beneficial for people with autism who can use portable devices to understand their own feelings and emotions and possibly adjust their social behaviour accordingly.

## **1.7 Organization of the Study**

This study consists of Seven chapters:

Chapter I: Introduction – This chapter is intended to contain the study's first sections, which include: Background of the study, Problem statement, Research objective, Research Questions, Significance of the study, Scope of the study.

Chapter II: Literature Review – This chapter concentrates on the preceding literature on About speech recognition through emotions.

Chapter III: methodology – The methodology and architecture utilized in this system to Construct a software solution are described in this chapter. It covers the system's hardware

And software requirements, as well as potential solutions for the research question,

Techniques, and concepts chosen.

Chapter IV: System analysis and design – This chapter discusses how the system works,

Which is the most important phase of our system. The following sections are also included:

Existing system, proposed system, System Requirements and the Feasibility of the study

Finally, this chapter describes the design of the system.

Chapter V: Implementation – This chapter confers about the design and development of

A system by using software tools and hardware devices. It also displays the most Important code which makes fundamental impact to the system functionality and Screenshot about system interface.

Chapter VI: Discussion this chapter will discuss the results of the research, and it provides the findings of these studies.

Chapter VII: Conclusion and Recommendation- This chapter concludes the research by

Giving the results of the research, challenges (limitations) and provides the necessary recommendations.



## **CHAPTER II: LITERATURE WORK**

### **2.0 introduction**

This chapter will go over several speech emotion recognition systems and how they relate to the speech emotion recognition system. Which include, machine learning voice recognition, emotion recognition from speech and other types of speech recognition will be discussed.

Speech Emotion Recognition system is defined as a collection of methodologies that use machine learning to process and classify speech signals in order to detect emotions. Such a system could be useful in areas such as interactive voice-based assistants or caller-agent conversation analysis.

### **2.1 History of machine learning**

The term machine learning was first coined in the 1950s when Artificial Intelligence pioneer Arthur Samuel built the first self-learning system for playing checkers. He noticed that the more the system played, the better it performed.

Fuelled by advances in statistics and computer science, as well as better datasets and the growth of neural networks, machine learning has truly taken off in recent years.

Today, whether you realize it or not, machine learning is everywhere – automated translation, image recognition, voice search technology, self-driving cars, and beyond.

### **2.2 What is machine learning?**

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to “self-learn” from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions. In short, machine learning algorithms and models learn through experience.

In traditional programming, a computer engineer writes a series of directions that instruct a computer how to transform input data into a desired output. Instructions are mostly based on an IF-THEN structure: when certain conditions are met, the program executes a specific action.

Machine learning, on the other hand, is an automated process that enables machines to solve problems with little or no human input, and take actions based on past observations.

## **2.3 Types of machine learning**

### **2.3.1 Supervised learning**

Gartner, a business consulting firm, predicts that supervised learning will remain the most utilized machine learning among enterprise information technology leaders in 2022. This type of machine learning feeds historical input and output data in machine learning algorithms, with processing in between each input/output pair that allows the algorithm to shift the model to create outputs as closely aligned with the desired result as possible. Common algorithms used during supervised learning include neural networks, decision trees, linear regression, and support vector machines.

This machine learning type got its name because the machine is “supervised” while it's learning, which means that you’re feeding the algorithm information to help it learn. The outcome you provide the machine is labelled data, and the rest of the information you give is used as input features.

For example, if you were trying to learn about the relationships between loan defaults and borrower information, you might provide the machine with 500 cases of customers who defaulted on their loans and another 500 who didn't. The labelled data “supervises” the machine to figure out the information you're looking for. Supervised learning is effective for a variety of business purposes, including sales forecasting, inventory optimization, and fraud detection. Some examples of use cases include:

Predicting real estate prices

Classifying whether bank transactions are fraudulent or not

Finding disease risk factors

Determining whether loan applicants are low-risk or high-risk

Predicting the failure of industrial equipment's mechanical parts

### **2.3.2 Unsupervised learning**

While supervised learning requires users to help the machine learn, unsupervised learning doesn't use the same labelled training sets and data. Instead, the machine looks for less obvious patterns in the data. This machine learning type is very helpful when you need to identify patterns and use data to make decisions. Common algorithms used in unsupervised learning include Hidden Markov models, k-means, hierarchical clustering, and Gaussian mixture models. Using the example from supervised learning, let's say you didn't know which customers did or didn't default on loans. Instead, you'd provide the machine with borrower information and it would look for patterns between borrowers before grouping them into several clusters.

This type of machine learning is widely used to create predictive models. Common applications also include clustering, which creates a model that groups objects together based on specific properties, and association, which identifies the rules existing between the clusters. A few example use cases include:

Creating customer groups based on purchase behaviour

Grouping inventory according to sales and/or manufacturing metrics

Pinpointing associations in customer data (for example, customers who buy a specific style of handbag might be interested in a specific style of shoe)

### **2.3.3 Reinforcement learning**

Reinforcement learning is the closest machine learning type to how humans learn. The algorithm or agent used learns by interacting with its environment and getting a positive or negative reward. Common algorithms include temporal difference, deep adversarial networks, and Q-learning. Going back to the bank loan customer example, you might use a reinforcement learning algorithm to look at customer information. If the algorithm classifies them as high-risk and they default, the algorithm gets a positive reward. If they don't default, the algorithm gets a negative reward. In the end, both instances help the machine learn by understanding both the problem and environment better. Gartner notes that most ML platforms don't have reinforcement learning capabilities because it requires higher computing power than most organizations have. Reinforcement learning is applicable in areas capable of being fully simulated that are either stationary or have large volumes of relevant data. Because this type of machine learning requires less management than supervised learning, it's viewed as easier to work with dealing with unlabelled data sets. Practical applications for this type of machine learning are still emerging. Some examples of uses include:

Teaching cars to park themselves and drive autonomously

Dynamically controlling traffic lights to reduce traffic jams

Training robots to learn policies using raw video images as input that they can use to replicate the actions they see

### **2.4 What is voice recognition (speaker recognition)?**

Voice or speaker recognition is the ability of a machine or program to receive and interpret dictation or to understand and perform spoken commands. Voice recognition has gained prominence and use with the rise of artificial intelligence and intelligent assistants, such as Amazon's Alexa and Apple's Siri

Voice recognition systems let consumers interact with technology simply by speaking to it, enabling hands-free requests, reminders and other simple tasks.

Voice recognition can identify and distinguish voices using automatic speech recognition (ASR) software programs. Some ASR programs require users first *train* the program to recognize their voice for a more accurate speech-to-text conversion. Voice recognition systems evaluate a voice's frequency, accent and flow of speech.

Although voice recognition and speech recognition are referred to interchangeably, they aren't the same, and a critical distinction must be made. Voice recognition identifies the speaker, whereas speech recognition evaluates what is said.

#### **2.4.1 How voice recognition works**

Voice recognition software on computers requires analog audio to be converted into digital signals, known as analog-to-digital (A/D) conversion. For a computer to decipher a signal, it must have a digital database of words or syllables as well as a quick process for comparing this data to signals. The speech patterns are stored on the hard drive and loaded into memory when the program is run. A comparator checks these stored patterns against the output of the A/D converter -- an action called pattern recognition. Speech is the basic way of interaction between the listener to the speaker by voice or expression. Humans can easily understand the speakers' message, but machines can't understand the speaker's word(Zhou et al., 2016).

Speech recognition is an intrinsically energetic procedure in speech model problems. Speech recognition gives the usual impression to believe the subject on recurrent neural networks (RNNs)(Kumar et al., 2022).

To correctly recognize the emotion from speech data, it is very important to extract the features which accurately represent the emotional aspect of speech signals. One of the biggest challenges in this field is to extract efficient features for the best classification of

emotions. Some notable works in this area include analysis and synthesis of emotional speech(Jeong-Sik ,Ji-Hwan , Yung-Hwan, 2009).

Existing studies have found that MFCCs are a far preferable way of analysing emotions compared to other commonly used speech features (e.g., loudness, formants, linear predictive coefficients etc.)(Hasan et al., 2004).

Speech emotion analysis is the process of identifying vocal cues in speech that serve as indicators of emotional state, mood, or tension. The key presumption is that it is possible to forecast a speaker's emotional state using objectively quantifiable clues. This idea is acceptable given that emotional emotions trigger physiological responses that have an impact on how we speak. For instance, the emotional state of dread typically causes muscle tension, a quick heartbeat, rapid breathing, and perspiration. The vibration of the vocal folds and the structure of the vocal tract alter as a result of these physiological actions. All of this has an impact on the vocal qualities of speech, which enable the listener to discern the emotional state that the speaker is experiencing (Gjoreski,Gjoresk, 2014).

There are various technologies and methods which are used to recognize and processing of speech which have their own methods and way of functioning to recognize the speech. The most popular approaches used in speech-related classifications in the following articles. In these, the trend-based and application-based publications are about vector machines for speech-based emotional recognition, age, gender, and speech-based speech recognition(Professor, Sreyas Institute of Engineering and Technology, Hyderabad, India. et al., 2020).

Recently, the combination of different kinds of features has been widely used for emotion recognition in speech (Bhavan et al., 2019) Showed that the combination of MFCCs, Mel-energy spectrum dynamic coefficients (MEDCs) and energy with a SVM classifier on a

self-constructed Chinese emotional database and the EmoDB. The difference between MEDCs and MFCCs is that MEDCs are calculated as the logarithmic average of energies after the filter bank, while MFCCs are calculated as the logarithmic after the filter bank (Chen et al., 2012) used a three-level speech emotion recognition model to solve the speaker independent emotion recognition problem and extracted the energy, zero crossing rate (ZCR), pitch, the first to third formants, spectrum centroid, spectrum cut-off frequency, correlation density, fractal dimension, and five Mel frequency bands energy. The three levels classify the six emotions pairwise, with each level providing finer classification than the last Schuler (Chen et al., 2012).

Proposed the usage of a multiple-stage classifier with a support vector machine over 7 emotional classes, with the aim of employing both acoustic and linguistic features for emotion classification. A deep belief network was used for spotting emotional key-phrases. Various classifiers (Gaussian Mixture Models (GMMs) SVMs, Neural Networks, Nearest Neighbours) were used for training on the acoustic features, and then combined with the belief network using a neural network and their performances evaluated(Liu et al., 2018).

Ensemble learning constitutes the process of combining the learning procedures of multiple models in order to give a final, (usually) stronger learner. Such methods have been used in a wide range of application areas — including credit scoring, medical diagnosis(Wang et al., 2011). Ensemble methods have been applied to audio data as well Schuller(Zareapoor & Shamsolmoali, 2015) presented an analysis of ensemble machine learning on speaker-independent speech emotion recognition, and reported improved accuracy on data scraped from movie content. Morrison (Morrison et al., 2007).

Particularly, bagged ensembles of support vector machines have been analyzed in a few works (Kim et al., 2002) used such an ensemble for the problem of fault detection in

rotating machinery. However, such works are few and far in between, and we hope to further analyze this model and apply it for emotion recognition from speech (Hu et al., 2007).

## **2.5 Sensory modalities for emotion expression**

There is vigorous debate about what exactly individual can express nonverbally. Humans can express their emotions through many different types of nonverbal communication including facial expressions, quality of speech produced, and physiological signals of the human body. In this section, we discuss each of these

### **2.5.1 Facial expressions**

The human face is extremely expressive, able to express countless emotions without saying a word. And unlike some forms of nonverbal communication, facial expressions are universal. The facial expressions for happiness, sadness, anger, surprise, fear, and disgust are the same across cultures.

### **2.5.2 Speech**

In addition to faces, voices are an important modality for emotional expression. Speech is a relevant communicational channel enriched with emotions: the voice in speech not only conveys a semantic message but also the information about the emotional state of the speaker. Some important voice feature vectors that have been chosen for research such as fundamental frequency, Mel-frequency cepstral coefficient (MFCC), prediction cepstral coefficient (LPCC), etc.

### **2.5.3 Physiological signals**

The physiological signals related to autonomic nervous system allow to assess objectively emotions. These include electroencephalogram (EEG), heart rate (HR), social media and Machine Learning electrocardiogram (ECG), respiration (RSP), blood pressure (BP), electromyogram (EMG), skin conductance (SC), blood volume pulse (BVP), and skin



temperature (ST). Using physiological signals to recognize emotions is also helpful to those people who suffer from physical or mental illness thus exhibit problems with facial expressions or tone of voice (Kerkeni et al., 2020).

## **2.6 Related Work**

In the last few years, new method is introduced where static feature vectors are obtained by using so called acoustic Low-Level Descriptors (LLDs) and descriptive statistical functionals, by using this approach a big number of large feature vectors is obtained. The downside is that not all of the feature vectors are of good value, especially not for emotion recognition. For that reason a feature selection method is often used(Gjoreski,Gjoresk, 2014).

Emotion is often entwined with temperament, mood, personality, motivation, and disposition. In psychology, emotion is frequently defined as a complex state of feeling that results in physical and psychological changes. These changes influence thought and behaviour. According to other theories, emotions are not causal forces but simply syndromes of components such as motivation, feeling, behaviour, and physiological changes(Kerkeni et al., 2020).

To be able to classify emotions using computer algorithms, we need to have a mathematical model describing them. The classical approach defined by psychologists is based on three measures that create a three-dimensional space that describes all the emotions, These measures or dimensions are pleasure, arousal, and dominance (Abbaschian et al., 2021) A combination of these qualities will create a vector that will be in one of the defined emotion territories, and based on that, we can report the most relevant emotion (Mehrabian, 1996) Using pleasure, arousal, and dominance, we can describe almost any emotion, but such a deterministic system will be very complex to implement for machine learning. Therefore, in machine learning studies, typically, we

use statistical models and cluster samples into one of the named qualitative emotions such as anger, happiness, sadness, and so forth. To be able to classify and cluster any of the mentioned emotions, we need to model them using features extracted from the speech; this is usually done by extracting different categories of prosody, voice quality, and spectral features (Gangamohan et al., 2016)

Any of these categories have benefits in classifying some emotions and weaknesses in detecting others. The prosody features usually focused on fundamental frequency (F0), speaking rate, duration, and intensity, are not able to confidently differentiate angry and happy emotions from each other(Gangamohan et al., 2016).

The immediate advantage that they have compared to prosody features is that they can confidently distinguish angry from happy. However, an area of concern is that the magnitude and shift of the formants for the same emotions vary across different vowels, and this would add more complexity to an emotion recognition system, and it needs to be speech content-aware(Vlasenko et al., 2011).

Enhancement of speech is used to improve speech performance through different measurements. The goal is to improve the understandability and overall sensory performance of distorted speech signals by using processing techniques related to speech algorithms, the main areas of speech enhancements include noise reduction and other applications, such as smartphones,



1Fig 2.1: List of Development in Technologies used for Speech Recognition

Fig 2.1: List of Development in Technologies used for Speech Recognition

Teleconference, recognition of speech and audio equipment. Three simple classes can group speech enhancement algorithms for noise reduction: spectral restoration, methods based on the model and filtering techniques (Professor, Sreyas Institute of Engineering and Technology, Hyderabad, India. et al., 2020).

In 2020, published a brief review on the importance of speech emotion datasets and features, noise reduction; ultimately, they analyze the significance of different classification approaches, including SVM and HMM. The strength of the research is the identification of several features related to speech emotion recognition; however, its weakness is the leak of more modern methods' investigation and briefly mentions convolutional and recurrent neural networks as a deep learning method(Basu et al., 2017).

The main question is: Are there any objective feature profiles of the voice that can be used for speaker emotion recognition? A lot of studies are done for the sake of providing

such feature profiles that can be used for representation of the emotions, but results are not always consistent. For some basic problems like distinguishing normal speech from angry speech or distinguishing normal speech from bored speech the experimental results converge, For example such converging results are showing that compared to normal speech, when expressing fear or happiness human speak with higher pitch (fundamental frequency) (Gjoreski,Gjoresk, 2014).

This simple analysis is just an example of how we can compare speech signals by using their physical characteristics. This simple approach cannot be used for speech emotion recognition. The problem arises when we have to distinguish emotional states like anger from happiness or fear from happiness. By using the basic speech audio features for describing these emotional states, the feature profiles are quite similar so distinguishing them is hard (Gjoreski,Gjoresk, 2014).

Recognition of emotions in audio signals has been a field of extensive study in the past. Previous work in this area included use of various classifiers like SVM, Neural Networks, Bayes Classifier etc. The number of emotions classified varied from study to study, they play an important aspect in evaluating the accuracy of the different classifiers. Reduction in the number of emotions used for recognition has generated more accurate results as depicted below. The following table summarises the previous study done on the topic.

Table 2.1: Summarises the previous study.

Study	Algorithm used	# emotion	#Accuracy (%)
[Kamran Soltani, Raja Noor Ainon,2007]	Two layer Neural Network	6	77.1

[Li Wern Chew, Kah Phooi Seng, Li-Minn Ang, Vish Ramakonar, 2011]	PCA, LDA and RBF	6 (divided into three independent classes)	81.67
[Taner Danisman, Adil Alpkocak, 2008]	SVM	4/5	77.5/66.8
[Lugger and Yang, 2007]	Bayes Classifier	6	74.4
[Yixiong Pan, Peipei Shen, Liping Shen, 2012]	SVM	3	95.1

To successfully implement a speech emotion recognition system, we need to define and model emotion carefully. However, there is no consensus about the definition of emotion, and it is still an open problem in psychology. According to Plutchik, more than ninety definitions of emotion were proposed in the twentieth century. Emotions are convoluted psychological states that are composed of several components such as personal experience, physiological, behavioural, and communicative reactions. Based on these definitions, two models have become common in speech emotion recognition: discrete emotional model, and dimensional emotional model(Akçay & Oğuz, 2020).

Automatic SER helps smart speakers and virtual assistants to understand their users better, especially when they recognize dubious meaning words. For example, the term “really” can be used to question a fact or emphasize and stress out a statement in both positive and negative ways. Read the following sentences in different ways: “I really liked

having that tool.” The same application can help translate from one language to another, especially as other languages have different ways of projecting emotions through speech. SER is also beneficial in online interactive tutorials and courses. Understanding the student’s emotional state will help the machine decide how to present the rest of the course contents. Speech emotion recognition can also be very instrumental in vehicles’ safety features. . It can recognize the driver’s state of mind and help prevent accidents and disasters. Another related application is in therapy sessions; by employing SER, therapists will understand their patients’ state and possibly underlying hidden emotions as well. It has been proven that in stressful and noisy environments like aircraft cockpits, the application of SER can significantly help to increase the performance of automatic speech recognition systems. The service industry and e-commerce can utilize speech emotion recognition in call centre’s to give early alerts to customer service and supervisors of the caller’s state of mind. In addition, speech emotion recognition has been suggested to be implemented in interactive movies to understand viewers’ emotions. The interactive film could then go along different routes and have different endings(Akçay & Oğuz, 2020).

To train machine learning algorithms to classify emotions, we need to have training datasets. For SER tasks, there are generally three types of training datasets, natural, semi natural, and simulated. The natural datasets are extracted from available videos and audios, either broadcasted on TV or online. There are also databases from call centers and similar environments. Semi-natural datasets are made by defining a scenario for professional voice actors and asking them to play them. The third and most widely used type, the simulated datasets, are similar to semi-naturals. The difference is that the voice actors are acting the same sentences with different emotions. Traditionally SER used to follow the steps of automatic speech recognition (ASR), and methods based on HMMS, GMMs, and SVMs were widespread. Those approaches needed lots of feature

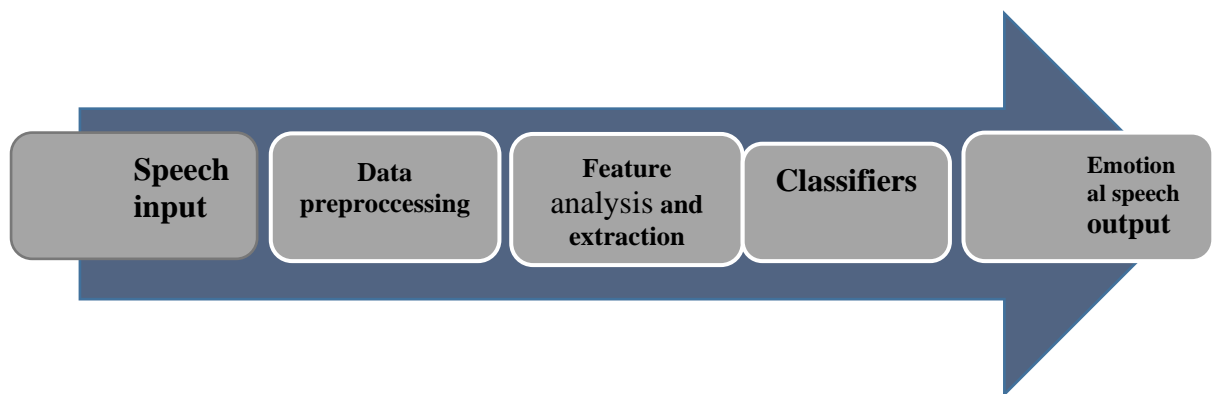
engineering and any changes in the features usually required restructuring the entire architecture of the method. However, lately, by the development of deep learning tools and processes, solutions for SER can be changed as well. There is a lot of effort and research on employing these algorithms to recognize emotions from the speech.

In addition to deep learning, more recently, along with improvements in recurrent neural networks and the use of long short-term memory (LSTM) networks, auto encoders, and generative adversarial models, there has been a wave of studies on SER using these techniques to solve the problem. The rest of the paper is organized as follows: In Section 2, we define SER, and in Section 3, we present some related studies. In Section 4, we provide a review of existing emotional speech datasets, and Section 5 is where we review several traditional and deep learning methods used in SER. Finally, in the last chapter, we discuss and conclude our work while proposing direction for future actions in SER (Akçay & Oğuz, 2020).

There are several emotional speech databases that are extensively used in the literature German, English, Japanese, Spanish, Chinese, Russian, Dutch etc. One main characteristic of an emotional speech database is the type of the emotions expressed in the speech. Whether they are simulated or they are extracted from real life situations. The advantage of having a simulated speech is that the researcher has a complete control over the emotion that it is expressed and complete control over the quality of the audio. However, the disadvantage is that there is loss in the level of naturalness and spontaneity. On the other hand, the non-simulated emotional databases consist of a speech that is extracted from real life scenarios like call-centers, interviews, meetings, movies, short videos and similar situations where the naturalness and spontaneity is kept. The disadvantage is that in these databases there is not a complete control over the expressed emotions. Also the low quality of the audio can be problem(Gjoreski,Gjoresk, 2014).

In general, there are two approaches in human emotions analysis. In the first approach the emotions are represented as discrete and distinct recognition classes. The other approach represents the emotional states in 2D or 3D space, where parameters like emotional distance, level of activeness, level of dominance and level of pleasure are be observed(Gjoreski,Gjoresk, 2014).

In this research we present a ML approach for automatic recognition of emotions from speech. Our approach uses the discrete type of approach; therefore the emotional states are represented by seven classes: anger, fear, sadness, happiness, boredom, disgust and neutral. Even though ML approaches have been proposed in the literature, our approach improves upon them by performing a thorough ML analysis, including methods for: feature extraction, feature standardization, feature selection, algorithm selection, and algorithm parameters optimization. With this analysis, we try to find the optimal ML configuration of: features, algorithms and parameters, for the task of emotion recognition in speech (Gjoreski,Gjoresk, 2014).



*2Figure 2.2: ML approach for emotion recognition.*

Figure 2.2: ML approach for emotion recognition.

Speech emotion recognition is a difficult problem for machine learning. The analysis of a sound signal is difficult to make as it includes various frequencies and features. Speech is digitized using signal processing methods and then sound characteristics are obtained through acoustic analysis. However, the overall success rate changes as the changes in



these characteristics differ according to the emotions (sadness, fear, anger, happiness, neutral, displeasure, etc.). Although different methods are utilized in both feature extraction and emotion recognition, the success rate varies according to emotions and databases(Sönmez,arol, 2019).

Sound cannot spread in space, because there are no particles, such as atoms or molecules, which could lead to the contraction and expansion of a substance. Sound travel also depends on the temperature of the environment. Sound also creates energy. The type of energy emerging from a substance's oscillation and vibration is called sound energy. Sound has physical qualities, such as frequency, wavelength, period, speed of travel, intensity, loudness, timbre, echo, pressure, amplitude, and resonance. Wavelength is the distance travelled by sound for the formation of the sound wave. The number of vibrations of sound in a unit of time (usually seconds) is called “frequency.” Its unit of measurement is Hertz (Hz). Unlike frequency, the period is the amount of time required for a full vibration(Sönmez, Rol, 2019).

Sound waves are grouped into three categories, based on their frequencies: the human ear can hear sounds with a frequency of between 20 Hz and 20000 Hz (20 kHz). The sound waves within the sensitivity thresholds of the human ear are called audible sound waves. These sounds can be produced in many ways such as using musical instruments or vocal cords. Sound waves less than 20 Hz are called infrasonic sound waves. For example, earthquake waves. Sound waves of more than 20000 Hz are called ultrasonic sound waves. Dogs and bats can hear sounds at this frequency. Sound waves are used in different ways in science and technology. These are Ultrasound devices are used when imaging internal organs, using ultrasonic (high frequency) waves. Sonars and echolocation are used in maritime for scanning fish, submarines, and geographical formations under the sea. Sound waves are also used to break up stones in the kidneys. Sound intensity is related to how loud or soft a sound is. Sound intensity depends upon the frequency of the

sound wave. We sense sound waves with a low frequency as soft, and sound waves with a high frequency, loud. Our perception of the highness of sound varies in proportion to the frequency of the vibration of particles in an environment. The crying sound of babies is loud, which means its frequency is high. On the other hand, an adult person's voice is soft, which means its frequency is low. The intensity of sound depends directly on the mechanical pressure on the eardrum sound depends on, and is directly proportional to, the energy and amplitude of sound waves. As the amplitude of sound waves increases, the energy and intensity of the sound increase. On the contrary, as the amplitude of sound waves decrease, the energy and intensity of the sound decline. The measure of sound intensity is called the sound level, and the unit of measurement of the sound level is a decibel (dB). Our ear can pick up sounds in between 0 and 140 dB(Sönmez,arol, 2019).

The technology of speech recognition and processing is evolving, increasing and improving. Technology for speech awareness can interact with other disabled persons. This allows the control of the digital system. Great opportunity in the future to extend the spoken network of engineering. Enhancing speech recognition can provide improved services to people with disabilities and provide our system with a secure environment with voice authentication. We are still well on the way before us because of the high level of competition on the market between this tech giant and the growing prevalence of companies jumping in to produce space content(Kim et al., 2002).

The Production and generation of Speech are recognized in four major steps. Activities that are related to speech, procurement, processing and generation of output and application of based on the function of the Speech extracted. Here, as vocal cords are used to produce speech from the mouth, sensors and recording techniques can detect it. The voice will then be processed using expression techniques and processes. Speech synthesis then takes place to rearrange the voice bits to a certain frequency to produce a coherent word that is further used by the product based on the use of speech technology But when

it comes to Speech Emotion recognition, the process is explained in five main steps and these steps are involved in each and every application related to SER which is based on machine learning techniques. They are classified as Speech input, Feature Extraction, Feature Selection, Classifier and Recognition of Emotion in Speech. They are:

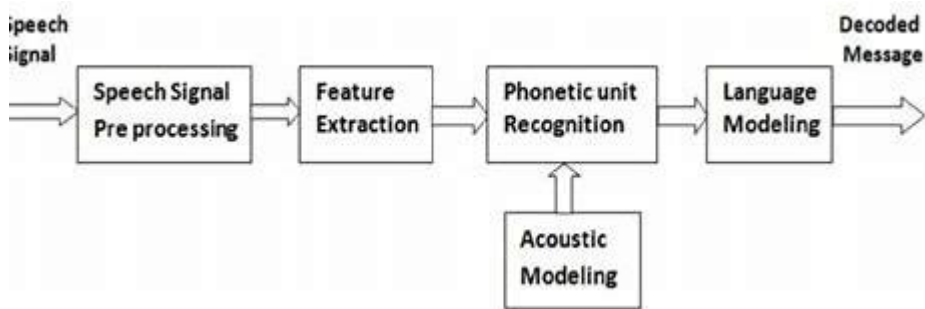


Figure 2.3: Process of Recognizing Emotion from Speech

Figure 2.3: Process of Recognizing Emotion from Speech

A typical set of emotion contains 300 Emotional States. While the primary emotions are Anger, Disgust, Fear, Joy, Sadness, and Surprise. In those, Anger, Joy, Sadness and fear are primary, these emotions are differentiated based on the corresponding changes that occur in speech rate, pitch, energy and spectrum one of the main speech features that indicate emotions is energy and study of energy. It depends on the short-term energy and short-term average amplitude. The speech input is taken based on the required format and then feature extraction is done by MFCC and LPCC methods and selection process based on the Artificial Neural Network (ANN) which are Machine Learning Techniques(Professor, Sreyas Institute of Engineering and Technology, Hyderabad, India. et al., 2020).

Let us see how the classification of Speech Processing is done base on each and every aspect and application by observing the following diagram. Speech processing is developing and increasing by data processing and providing traditional transportation for

many workplaces that integrate application-based computing and telephony through different platforms. The present device becomes a terminal for the personal remote workstation with added speech input and output capabilities, allowing access to its functionality from anywhere depending on its result and outcome. The required speech is also minimized by the advent of effective and spectral algorithms, very large circuits with integrated and processors of digital signals and detection of the audio signal based on these efficient algorithms and methods. Speech recognition is typically used to translate the spoken word to a specific speech message response, while speech verification is used to check the voice features of the clients. The aim of speech recognition systems is simply to understand the speaker's spoken word and develop the speaker's identity(Professor, Sreyas Institute of Engineering and Technology, Hyderabad, India. et al., 2020).

In contributions to the other published surveys, this research provides a thorough study of significant databases and deep learning discrete approaches in SER. The reason for not focusing on the other older techniques is recent progress in neural networks and, more specifically, deep learning. Based on the best of our knowledge, this study is the first survey in SER focusing on deep learning along with unified experimental results that proposes approaches to enhance the available methods' results(Abbaschian et al., 2021).

## **2.7 conclusion**

Speech is the basic way of interaction between the listener to the speaker by voice or expression. Humans can easily understand the speakers' message, but machines can't understand the speaker's word

In this paper, discriminated speaking technology are spotlighted on the feature extraction, improvement, segmentation and progression of speech emotion recognition. Initially, the trained RNN layer-based feature extraction is done to get the speech signal's high-level features.

## **CHAPTER III: RESEARCH METHODOLOGY**

### **3.0 Introduction**

This chapter includes the study's presentation as well as the research methodologies used to complete it. There's also a system description, a system overview, system features, and system requirements.

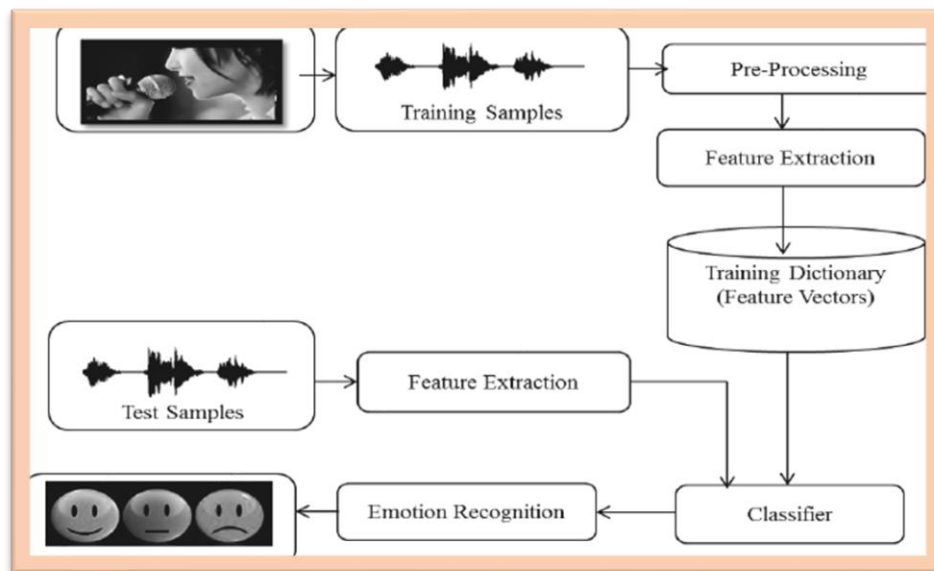
First an emotional speech database is used, which consists of simulated and annotated utterances. Next, feature extraction is performed by using open-source feature extractor. Then, feature selection method is used for decreasing the number of features and selecting only the most relevant ones. Finally, the emotion recognition is performed by a classification algorithm.

### **3.1 System Description**

The performance of an SER system using machine learning is evaluated using various metrics such as accuracy, precision, recall, and F1 score. This system is typically trained and tested using speech data, where each speech sample is labeled with the corresponding emotional state. Overall, an SER system using machine learning is a powerful tool for recognizing and analyzing the emotional content of speech. With the increasing availability of speech data and advances in machine learning techniques, SER systems are expected to become more accurate and robust in the future.

### **3.2 System architecture**

The system architecture of a speech emotion recognition system using machine learning involves several stages, including signal pre-processing, feature extraction, feature selection, classification, and system evaluation. The success of the system depends on the choice of features, the machine learning algorithm used, and the quality of the TESS speech dataset used for training and testing the system.



4Figure 3.1 System Architecture

Figure 3.1 System Architecture

### 3.3 System features

This system predicts emotion of speech. Our system will have a variety of features, including:

**Speech signal sampling** is the reduction of a continuous-time signal to a discrete-time signal. A common example is the conversion of a sound wave to a sequence of "samples".

**Pre-processing Data** the data pre-processing can often have a significant impact a supervised ML algorithm's extension ability. One of the most challenging tasks in deductive ML is the reduction of background occurrences. Is a program that processes its input data to produce output that is used as input to another program like a compiler.

**Feature extraction** is the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. , many common features are extracted, such as energy, pitch, formant, and some spectrum features such as linear prediction coefficients (LPC), Mel-

frequency cepstrum coefficients (MFCC), and modulation spectral features. In this work, we have selected modulation spectral features and MFCC, to extract the emotional features.

**Feature Selection** is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context to deal with this issue, we used a method for feature selection. Feature selection (FS) aims to choose a subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to higher recognition accuracy.

Recursive feature elimination (RFE) uses a model (e.g., linear regression or SVM) to select either the best- or worst-performing feature and then excludes this feature.

**Classification** Many machine learning algorithms have been used for discrete emotion classification. The goal of these algorithms is to learn from the training samples and then use this learning to classify new observation. In fact, there is no definitive answer to the choice of the learning algorithm; every technique has its own advantages and limitations.

For this reason, here we chose to compare the performance of three different classifiers:

**Support vector machines (SVM)** are an optimal margin classifier in machine Learning. It is also used extensively in many studies that related to audio emotion recognition. It can have a very good classification performance compared to other classifiers especially for limited training data.

**Multivariate linear regression classification (MLR)** is a simple and efficient computation of machine learning algorithms, and it can be used for both regression and classification problems.

- ✓ **Training a classifier** is where you determine the optimal values (based on your training content set) to make the trade-offs that make sense to your application.
- ✓ **Test Train Split** When machine learning algorithms are used to generate predictions on data that was not used to train the model, this approach is used to estimate their performance.

### **3.4 System requirement**

System Requirement Specification Our system will have a lot of integrated requirements, including hardware and software. The software requirement of our system python & JupyterNotebook.

#### **3.4.1 Hardware Component**

The hardware component of our system will consist of the following components:

- ✓ Computer Hp
- ✓ Cori5
- ✓ Ram8
- ✓ SSD
- ✓ Speaker
- ✓ Microphone

#### **3.4.2. Software component**

The following components will make up the software component of our system:

- ✓ Operating system(windows, Linux )
- ✓ python



- ✓ Jupyter Notebook
- ✓ Python flask
- ✓ Visual Code (Vs Code).

**Python** is an interpreter, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and 37 therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

**Python flask** is Flask is a popular Python web framework, meaning it is a third-party Python library used for developing web applications. Also is a small and lightweight Python web framework that provides useful tools and features making

creating web applications in Python easier, It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around Werkzeug and Jinja and has become one of the most popular Python web application frameworks. Flask offers suggestions, but doesn't enforce any dependencies or project layout.

**A Jupiter Notebook** is an open-source web application that allows data scientists to create and share documents that include live code, equations, and other multimedia resources. Jupyter notebooks are used for all sorts of data science tasks such as exploratory data analysis (EDA), data cleaning and transformation, data visualization, statistical modeling, machine learning, and deep learning. Jupyter notebooks are especially useful for "showing the work" that your data team has done through a combination of code, markdown, links, and images. They are easy to use and can be run cell by cell to better understand what the code does.

**Visual Studio Code** comes includes an incredibly quick source code editor that is perfect for daily use. With support for hundreds of languages, VS Code's syntax highlighting, bracket-matching, auto indentation, box-selection, and snippets enable you to be more productive more quickly.

## **CHAPTER IV SYSTEM ANALYSIS AND DESIGNS**

### **4.1 Introduction**

This chapter focuses on the analysis and design of our research on emotion-based speech recognition using machine learning, which is one of the key forms of communication among humans. In the system analysis, we will address the existing method of speech emotion, which is the manual method, the disadvantages of the current method, and the necessity for this system, as well as the requirements of our system. The system requirements can be divided into two categories. Functional and non-functional requirements; functional requirements explain the necessary functions of our system that are unique to the system; non-functional requirements discuss the system's common functions. This chapter's main focus is on system design.

### **4.2 System Analysis**

System analysis involves the study of machine learning methods of speech emotion prediction and the most important face facts is to collect dataset after teaching the models by removing any Data cleaning as mentioned Chapter three after those models have been tested and any data cleaning will be done, the interface will be streamlined for users to use this system.

### **4.3 System current**

Existing systems have limitations in accurately recognizing and distinguishing between various emotions conveyed in speech, some current systems have limited integration capabilities or may not be easily accessible to users

#### **4.3.1 Current drawbacks**

- **Limited Emotion Recognition:** Existing systems may have limitations in accurately recognizing and distinguishing between various emotions conveyed in speech.

- **Lack of Flexibility:** Existing systems may not support audio inputs of different extensions or may have restrictions on the types of audio files they can process.
- **Lack of Real-Time Processing:** Some existing systems may suffer from delays or latency in processing audio inputs and providing emotion predictions.
- **Inadequate Training Data:** Many existing systems may not have access to a comprehensive and diverse dataset for training the emotion recognition model.

#### **4.4 The Need of This Work**

An SER is an important for various reasons. This system can increase and improve the communication. SER can interact with other disabled persons. also It can recognize the driver's state of mind and help prevent accidents and disasters. Another related application is in therapy sessions; by employing SER, therapists will understand their patients' state and possibly underlying hidden emotions as well.

#### **4.5 System proposed**

To deal with the problem, we developed automatic emotion prediction using machine learning techniques, that takes all voice extensions like mp3 or .wav formats and returns a result that shows by emoji's with text of the voice. So, the users can analyze and understand our system. This system is designed to prevent the approval of emotion prediction. Our Speech Emotion Recognition system aims to provide an improved and reliable solution for accurately recognizing and analysing emotions in speech. Our system is built using Flask, a web framework, allowing seamless integration into various applications and platforms, making it accessible to a wide range of users.

#### **4.6 System requirements**

In this section, researchers discuss requirements for the speech emotion System and classify the requirement into two parts Functional and Non-functional requirements

#### **4.6.1 Functional Requirements**

Functional Requirements is requirements that describe an activity of your system.

The following are the requirements that are proposed system should functionally attain:

- ❖ Audio Input: The system should be able to accept audio inputs of various extensions, such as WAV, MP3, or OGG
- ❖ Emotion Prediction: The system should accurately predict one of the seven emotions (Angry, Happy, Sad, Surprise, Disgust, Fear, and Normal) based on the input audio.
- ❖ Training and Testing: The system should be capable of training a machine learning model using the provided dataset and evaluating its performance. It should be able to update and retrain the model as new data becomes available.
- ❖ Real-Time Processing: The system should process the input audio and provide emotion predictions in real-time or within an acceptable timeframe.
- ❖ User Interface: The system should have a user-friendly interface where users can interact with the system, upload audio files, and receive emotion predictions as output.

#### **4.6.2 Non-Functional Requirements**

The following are the requirements that the proposed system should non-functionally achieve:

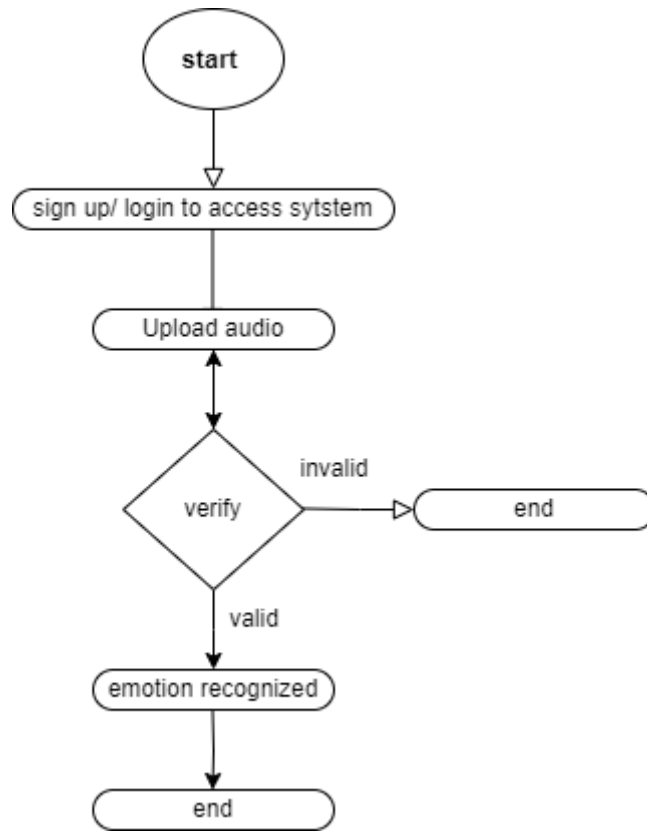
- ❖ Accuracy: The system should achieve a high level of accuracy in predicting emotions from audio inputs, ensuring reliable results.
- ❖ Security: The system should implement appropriate security measures to protect user data and prevent unauthorized access.
- ❖ Usability: The system should be designed with a user-centric approach, ensuring a pleasant and intuitive user experience.

- ❖ **Ethical Considerations:** The system should be designed and trained with ethical considerations in mind, avoiding biases and promoting fairness in emotion recognition.
- ❖ **Performance:** The system should be able to handle multiple concurrent user requests efficiently, providing quick responses for emotion prediction.

#### **4.7 System design**

System design is the process of creating the architecture, modules, and components of a system, as well as the various interfaces between those components and the data that passes through it. The aim of system design is to make a new system as good as the old one. Use case diagram, Flow chart diagram or UML are used and analyzed to display the design of the system, including the following Designs.

#### **Flow Chart Diagram**



5Figure 4.1 follow chart diagram

Figure 4.1 follow chart diagram

#### 4.7.1 System Dataset

For SER tasks, there are generally three types of training datasets, natural, semi-natural, and simulated. The natural datasets are extracted from available videos and audios, either broadcasted on TV or online. There are also databases from call centers and similar environments. The performance and robustness of the recognition systems will be easily affected if it is not well trained with a suitable database. Therefore, it is essential to have sufficient and suitable phrases in the database to train the emotion recognition system and subsequently evaluate its performance. There are three main types of databases: acted emotions, natural spontaneous emotions, and elicited emotions. In this work, we used an acted emotion databases because they contain strong emotional expressions. The literature on speech emotion recognition shows that the majority of

studies have been conducted with emotional acted speech. In this section we detailed the emotional speech database used for classifying discrete emotions in our experiments on emotion recognition with raw audio, we built a dataset from Toronto Emotional Speech Set (TESS). Is an acted dataset primarily developed for analysing the effect of age on the ability to recognize emotions this dataset is all comprised of two female actors, about 60 and 20 years old? Each actor has simulated seven emotions for 200 neutral sentences. Emotions in this dataset are: angry, pleasantly surprised, disgusted, happy, sad, fearful, and neutral. To label the dataset, 56 undergraduate students were asked to identify emotions from the sentences. After the identification task, the sentences with over 66% confidence have been selected to be in the dataset.



## **CHAPTER V: IMPLEMENTATION AND TESTING**

### **5.1 introduction**

In this chapter, we discuss the details of our proposed SER system's implementation and testing. This chapter also includes Snapshots of the project's key components as well as descriptions of how the system operates. This project explains how to build a SER using a neural networking (NN) model that continuously predicts emotions. We also present the results of our experiments and analyze the performance of the proposed system.

### **5.2 Overview of the implementation environment**

To implement our proposed SER system, we have chosen to use two different environments Kaggle and jupyter notebooks and python flask. These platforms provide a web-based interface for running Jupyter notebooks and executing code, with access to hardware resources such as GPUs and TPUs, which are essential for training deep learning models like the one used in our SER system.

Kaggle is a platform for data science competitions and collaborative data science projects. Kaggle provides a free cloud-based environment for running Jupyter notebooks, with access to GPUs and TPUs. In addition to the computing resources, Kaggle also provides access to a large community of data scientists and machine learning experts, making it a valuable resource for learning and collaboration up to 16 hours at a time, making it an excellent choice for experimenting with deep learning models without the need for expensive hardware. Kaggle allows users to collaborate with other users, find and publish datasets.

In our implementation of the SER system, we used Kaggle to train and evaluate our model. We leveraged the GPU and TPU resources available on this platform to accelerate the training process and achieve better performance.

### 5.3 Data Collection & Understanding Process

We collecting Audio data from publicly available and combining it with self-made audio samples. The publicly available dataset is taken from kaggle included 2800 audios appraisal record; this data is in .wav file format. System will read this file using “pandas” which is python library.

#### Important libraries

```
import pandas as pd
import numpy as np
import os
import seaborn as sns
import matplotlib.pyplot as plt
import librosa
import librosa.display
from IPython.display import Audio
import warnings
warnings.filterwarnings('ignore')
```

6Figure 5.1 libraries

Figure 5.1 libraries

```
## Create a dataframe
df = pd.DataFrame()
df['speech'] = paths
df['label'] = labels
df.head()
```

	speech	label
0	/kaggle/input/toronto-emotional-speech-set-tes...	fear
1	/kaggle/input/toronto-emotional-speech-set-tes...	fear
2	/kaggle/input/toronto-emotional-speech-set-tes...	fear
3	/kaggle/input/toronto-emotional-speech-set-tes...	fear
4	/kaggle/input/toronto-emotional-speech-set-tes...	fear

7Figure 5.2 Data collection

Figure 5.2 Data collection

## 5.4 Data pre-processing

Pre-processing is an essential part of audio classification. Transforming audio into something understandable to an algorithm is crucial to avoid misleading predictions by the model.

## 5.5 Training and Test Data

To pre-process the audio data for our SER system, we converted the audio recordings to Wave show and spectrograms using a Librosa library .we then extracted the data and used Mel-frequency cepstral coefficients (MFCC) It extract 2800 audio files and 40 features of audio file. We then convert the feature into 1 dimension.

To train our SER model, we split the pre-processed dataset into batches for training and validation. We implemented batch training and validation sets to improve the training efficiency of the model.

We trained the model for 50 epochs with a learning rate of 0.2 and a batch size of 64. We split the data using 70/30 ration meaning 70% of the data is used to train the model, and the remaining part, 30%, is used to test the model. We trained the model, using model selection, with the data samples from the training set.

Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 256)	264192
dropout (Dropout)	(None, 256)	0
dense (Dense)	(None, 128)	32896
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256
dropout_2 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 7)	455

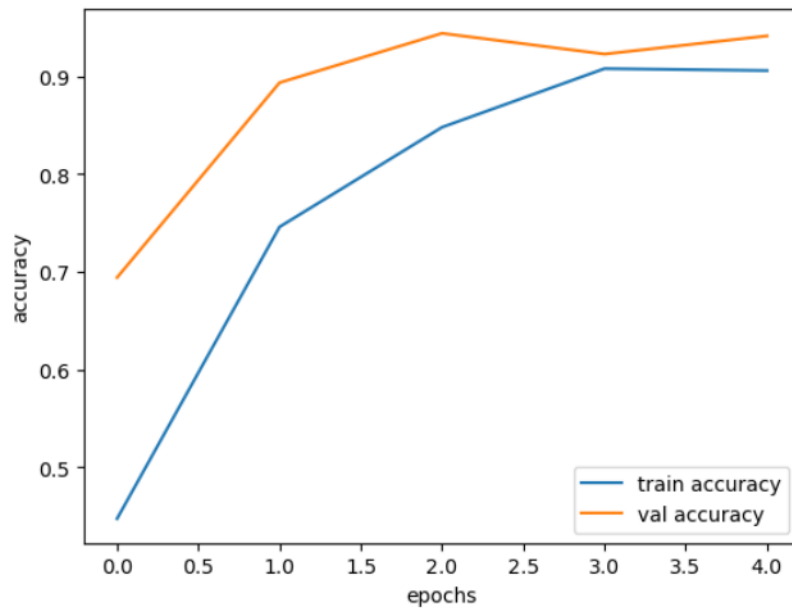
=====

Total params: 305,799  
Trainable params: 305,799  
Non-trainable params: 0

8Figure 5.3 Model Summary

Figure 5.3 Model Summary

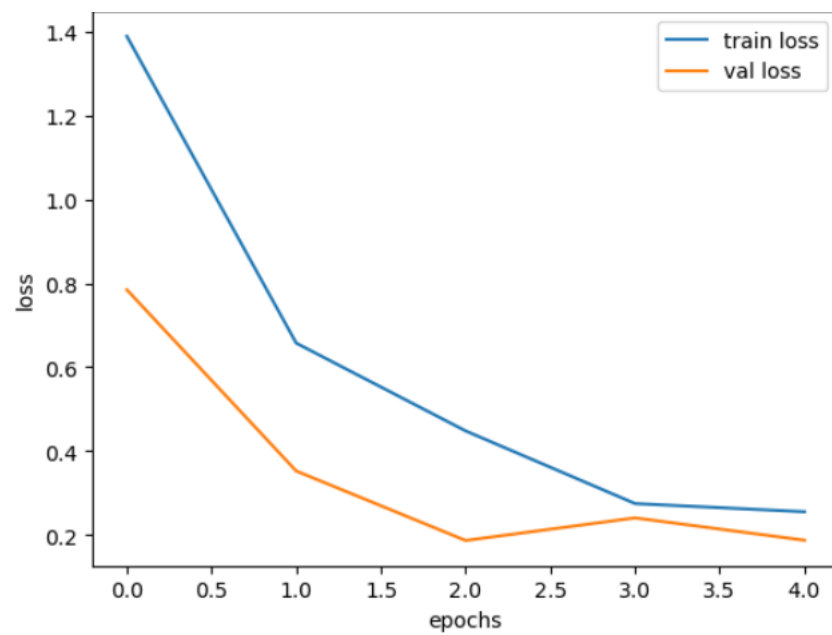
Display validation and Train Accuracy graph



9Figure 5.4 train accuracy and Val accuracy graph

Figure 5.4 train accuracy and Val accuracy graph

Display the loss and validation loss graph



10Figure 5.5 train loss and Val loss

Figure 5.5 train loss and Val loss

## 5.6 Classification Algorithms Accuracy Rate

Table 5.1 Accuracy Rate

Algorithm	Prediction Accuracy
Artificial Neural networking	98.73%

## 5.6 Model Deployment

After we have trained our SER model, we will deploy it in a production environment where it will be used for real-time speech recognition. To accomplish this, we used python Flask, an open-source platform for deploying and managing machine learning models.

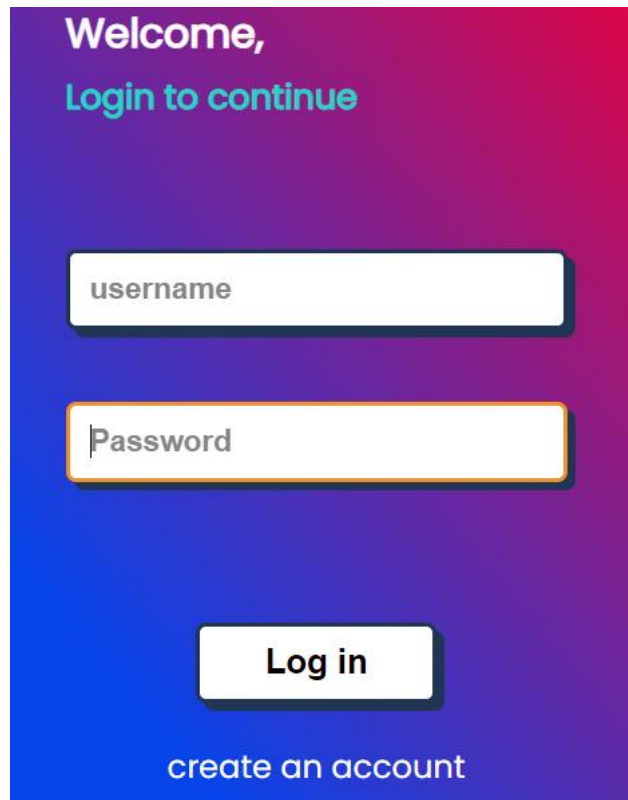
Flask is a popular Python web framework, meaning it is a third-party Python library used for developing web applications. We used Flask to create a RESTful API that can receive audio files and return the corresponding emotions using our SER model.

To connect the Flask server to a React client, we created a frontend web application using html/CSS, and python for backend the client sends HTTP requests to the Flask server through the REST API, and displays the transcriptions on the user interface.

## 5.7 System Snapshots

We took snapshots some of the most significant parts of the Administrators & Merchants, and these Snapshots will help system users to understand it effectively.

### 5.7.1 Login page



Welcome,  
Login to continue

username

Password

Log in

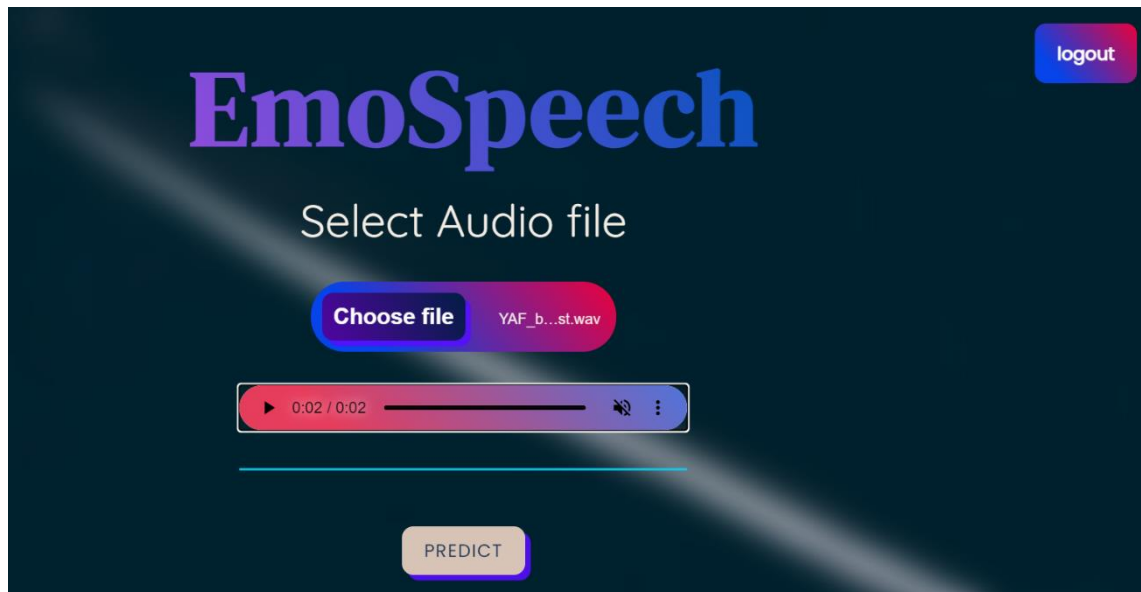
[create an account](#)

*11Figure 5.6 Login Page*

Figure 5.6 Login Page

The first page that shows when the portal loads is the Login Page, where the user must enter a valid username and password to determine, the login form is used to validate the user whether he or she is a system approved user

### 5.7.2 Dashboard page



12 Figure 5.7choose file/ upload audio

Figure 5.7choose file/ upload audio

After a user uploads or selects an audio file, they will be redirected to the page shown in Figure 5.8. This page provides the user with several options, including the ability to listen to the audio, to see audio as a wave, to download the audio, and to change playback speed. Additionally, the page also includes a section where the user can make a prediction to get results, logout the page.

### 5.7.8 Result Page

Show the result of prediction page where the users can view emotions of the audio selected and users can get results by emotion and text.



13

Figure 5.8 result page

Figure 5.8 result page



## CHAPTER VI: DISCUSSION

### 6.1 Introduction

In this chapter, we will review several studies that have addressed the challenge of developing SER systems. We will discuss the findings of these studies, including the data augmentation, and deep learning methods in improving SER performance for these languages. We will discuss the findings of these studies, including the importance of language modelling, data augmentation, and deep learning methods in improving SER performance for these languages. Finally, we will discuss the challenges that remain in developing SER systems for complex emotions and the need for further research to address these challenges.

### 6.2 Results

The results of our study indicate that the development of an SER system is feasible and holds great potential for improving the accessibility of emotion recognition technology and it is recreational with good accuracy. Through our analysis of an audio data in the form of numerical, we were able to identify the emotion of different audio samples and that are essential for the accurate recognition of emotions from the speech.

To assess the performance of the introduced RNN-based SER architectures, we performed speaker-independent SER experiments using Toronto Emotional Speech Set (TESS). Is an acted dataset primarily developed for analysing the effect of age on the ability to recognize emotions this dataset is all comprised of two female actors, about 60 and 20 years old? Each actor has simulated seven emotions for 200 neutral sentences. Emotions in this dataset are: angry, pleasantly surprised, disgusted, happy, sad, fearful, and neutral. In this section, experimentation results are presented and discussed. We report the recognition accuracy of using LSTM network (RNN) classifiers. Experimental evaluation is performed on the Toronto databases. The neural network structure used is a

simple LSTM. It consists of two consecutive LSTM with 256 layers with dense 128 activation that gets 40 features as input, followed by two classification dense layers. Features from data are scaled to [40, 1] before applying classifiers. Scaling features before recognition is important, because when a learning phase is fit on unscaled data, it is possible for large inputs to slow down the learning and convergence and in some cases prevent the used classifier from effectively learning for the classification problem.

As a baseline SER system, we used a RNNS classifier with long short-term memory (LSTM) its relative insensitivity to gap length is its advantage over other RNNs, hidden Markov models and other sequence learning methods. It aims to provide a short-term memory for RNN It is applicable to classification, processing and predicting data based on time series, such as speech recognition.

Our findings suggest that the proposed deep learning model with RNNs, holds great potential for improving the accessibility of speech emotion recognition technology.

### **6.3 limitations**

Our study has several limitations that should be considered when interpreting the results, there are many advancements on speech emotion recognition systems, and there are still several obstacles that need to be removed for successful recognition. One of the most important limitations is the generation of the Somali dataset that is used for the learning process. Most of the data sets used for SER are acted or elicited that are recorded in special silent rooms. However, the real-life data is noisy and has far more different characteristics than the others. Although natural data sets are also available, they are fewer in numbers. There are legal and ethical problems to record and use natural emotions. Most of the utterances in natural data sets are taken from talk-shows, call-center recordings, and similar cases where the involved parties are informed of the recording. Somali data sets do not contain all emotions and may not reflect the emotions that are felt. In addition, there are problems during the labelling of the utterances. There are human annotators

labelling the speech data after the utterances are recorded, and there is no annotators in Somalia. The actual emotion felt by the speaker and emotions perceived by human annotators may show differences. Even the recognition rates of human annotator are not over 90%. In Favor of humans, however, we believe that we also depend on the content and the context of the speech as we are evaluating. There are also cultural and language effects on SER. There are several studies available working on cross-language SER. However, the results show that current systems and features used are not sufficient for it. The intonation of emotions on speech among various languages may show differences for example. An overlooked challenge is the case of multiple speech signals, where the SER system has to decide which signal to focus on. Although it can be handled via a speech separation algorithm in the preprocessing stage, current systems fail to notice this problem.

In most modern SER systems, semi-natural and simulated datasets are utilized that are acted in nature, not noisy, and far from reality.

## **CHAPTER VII: CONCLUSION AND FUTURE WORK**

### **7.1 introduction**

The goal of this chapter is to provide a clear and succinct summary of the research's findings, which were reached after six months of in-depth inquiry and analysis. The chapter discusses a variety of significant topics, including a detailed explanation of the research's conclusion and an analysis of the extent to which the research objectives outlined in chapter one were met. Furthermore, the chapter offers useful guidelines and recommendations for people who may be interested in performing similar research in the future, as well as suggestions for areas of further analysis and exploration.

### **7.2 Conclusion**

Our study, we presented speech emotion recognition (SER) system using machine learning algorithms (RNN) to classify seven emotions. Thus, two types of features (MFCC and MS) were extracted from acted databases (TESS databases), and a combination of these features was presented. In fact, we study how classifiers and features impact recognition accuracy of emotions in speech. A subset of highly discriminant features is selected. Feature selection techniques show that more information is not always good in machine learning applications. The machine learning models were trained and evaluated to recognize emotional states from these features. SER reported the best recognition accuracy rate of 98% is achieved by RNN classifier without SN and with FS, on the TESS database using RNN classifier. Overall, our research project highlights the potential of deep learning-based SER models, providing a valuable contribution to this field that can benefit communities worldwide.

### **7.3 Recommendation for Future work**

It is recommended to support the used dataset with a greater number of algorithms to get high accuracy for the predictive model the study speech emotion recognition to improve

the current algorithms it can be concluded that. There are many factors that need to be considered to predict with more accuracy.

- To approach multiple algorithms will improve the accuracy and correctness of SER such as MLR, SVM, means algorithm can be applied to emotion prediction.

## REFERENCES

- Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. S. (2021). *Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models*.
- Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication, 116*, 56–76. <https://doi.org/10.1016/j.specom.2019.12.001>
- Basu, S., Chakraborty, J., Bag, A., & Aftabuddin, M. (2017). *A review on emotion recognition using speech*. 109–114.
- Bhavan, A., Chauhan, P., Hitkul, & Shah, R. R. (2019). Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems, 184*, 104886. <https://doi.org/10.1016/j.knosys.2019.104886>
- Chen, L., Mao, X., Xue, Y., & Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing, 22*(6), 1154–1160.
- Gangamohan, P., Kadiri, S. R., & Yegnanarayana, B. (2016). Analysis of emotional speech—A review. *Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions*, 205–238.
- Gjoreski, Gjoresk, M., Hristijan i. (2014). *Machine Learning Approach for Emotion Recognition in Speec*.
- Hasan, M. R., Jamil, M., & Rahman, M. (2004). Speaker identification using mel frequency cepstral coefficients. *Variations, 1*(4), 565–568.
- Hu, Q., He, Z., Zhang, Z., & Zi, Y. (2007). Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble. *Mechanical Systems and Signal Processing, 21*(2), 688–705. <https://doi.org/10.1016/j.ymssp.2006.01.007>

Jeong-Sik ,Ji-Hwan , Yung-Hwan, P., Kim,Oh. (2009). *Feature vector classification based speech emotion recognition for service robots.*

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. In A. Cano (Ed.), *Social Media and Machine Learning*. IntechOpen. <https://doi.org/10.5772/intechopen.84856>

Kim, H.-C., Pang, S., Je, H.-M., Kim, D., & Bang, S.-Y. (2002). Support Vector Machine Ensemble with Bagging. In S.-W. Lee & A. Verri (Eds.), *Pattern Recognition with Support Vector Machines* (Vol. 2388, pp. 397–408). Springer Berlin Heidelberg. [https://doi.org/10.1007/3-540-45665-1\\_31](https://doi.org/10.1007/3-540-45665-1_31)

Kumar, Dr. T., Villalba-Condori, K. O., Arias-Chavez, D., K., R., M, K. C., & S., Dr. S. R. (2022). An Evaluation on Speech Recognition Technology based on Machine Learning. *Webology*, 19(1), 646–663. <https://doi.org/10.14704/WEB/V19I1/WEB19046>

Liu, Z.-T., Wu, M., Cao, W.-H., Mao, J.-W., Xu, J.-P., & Tan, G.-Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273, 271–280.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14, 261–292.

Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2), 98–112.

Professor, Sreyas Institute of Engineering and Technology, Hyderabad, India., Kumar\*, S., Jason, C. A., & M.Tech Student, Sreyas Institute of Engineering and Technology, Hyderabad, India. (2020). An Appraisal on Speech and Emotion Recognition

- Technologies based on Machine Learning. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(5), 2266–2276. <https://doi.org/10.35940/ijrte.E5715.018520>
- Sönmez,arol, Y. Ü., Asaf. (2019). *New Trends in Speech Emotion Recognition*.
- Vlasenko, B., Philippou-Hübner, D., Prylipko, D., Böck, R., Siegert, I., & Wendemuth, A. (2011). *Vowels formants analysis allows straightforward detection of high arousal emotions*. 1–6.
- Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1), 223–230. <https://doi.org/10.1016/j.eswa.2010.06.048>
- Zareapoor, M., & Shamsolmoali, P. (2015). Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. *Procedia Computer Science*, 48, 679–685. <https://doi.org/10.1016/j.procs.2015.04.201>
- Zhou, N.-R., Li, J.-F., Yu, Z.-B., Gong, L.-H., & Farouk, A. (2016). New quantum dialogue protocol based on continuous-variable two-mode squeezed vacuum states. *Quantum Information Processing*, 16(1), 4. <https://doi.org/10.1007/s11128-016-1461-2>
- Bhavan, A., Chauhan,P. ,Shah,R. R. (2019). *Bagged support vector machines for emotion recognition from speech ☆*.
- Gjoreski, M. ,Gjoreski,H. ,KulakovA. (2014). *Machine Learning Approach for Emotion Recognition in Speech*.
- Kerkeni, L., Serrestou,Y. ,Mbarki,M. ,A. ,Cleder,C. (2019). *Automatic Speech Emotion Recognition Using Machine Learning*.



## Appendix A: Importing libraries to initialize flask

```
import pandas as pd

import numpy as np

from tensorflow import keras

from flask import Flask, render_template, request, redirect, url_for

import librosa

import librosa.display

# loading the Trained Model

model = keras.models.load_model('./model/speech_model.h5')

# Categories (Emotion category)

labels = ['angry', 'fear', 'happy', 'neutral', 'ps', 'sad']

# Initializing Flask

app = Flask(__name__, template_folder='templates', static_folder='staticFiles')

# This function is loading the audio and extracting the data out of it

def extract_mfcc(filename):

    y, sr = librosa.load(filename, duration=3, offset=0.5)

    mfcc = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=40).T, axis=0)

    return mfcc
```

## Appendix B: login and singup

```
# to handel username and password for Login and Signup

def write_to_file(username, password):

    with open('users.txt', 'a') as file:

        file.write(f'{username}:{password}\n')

# to check if user already exist

def check_credentials(username, password):

    with open('users.txt', 'r') as file:

        for line in file:

            stored_username, stored_password = line.strip().split(':')

            if username == stored_username and password == stored_password:

                return True

        return False

# Home page rendering

@app.route("/home")

def home():

    return render_template("home.html")

@app.route("/about")
```

```

def about():

    return render_template("about.html")

# Signup Page rendering

@app.route('/signup', methods=['GET', 'POST'])

def signup():

    if request.method == 'POST':

        username = request.form['username']

        password = request.form['password']

        write_to_file(username, password)

        return redirect(url_for('login'))

    return render_template('signup.html')

# Login page rendering

@app.route('/', methods=['GET', 'POST'])

def login():

    if request.method == 'POST':

        username = request.form['username']

        password = request.form['password']

        if check_credentials(username, password):

            return redirect(url_for('home'))

        else:

```

```

        return 'Invalid username or password'

    return render_template('login.html')

# This is used to handel prediction

@app.route('/predict2', methods=['GET', 'POST'])

def upload_file1():

    if request.method == 'POST':

        f = request.files['file']

        print("filename ",f.filename)

        paths = [f.filename]

        pred_df = pd.DataFrame({'speech':paths})

        pred_X_mfcc = pred_df['speech'].apply(lambda x: extract_mfcc(x))

        pred_X = [x for x in pred_X_mfcc]

        pred_X = np.array(pred_X)

        pred_X = np.expand_dims(pred_X, -1)

        pred = labels[model.predict(pred_X)[0].argmax()] # this is used to predict the
audio emotion

        return render_template('reaction.html',mode=str(pred)) # sending the prediction
value to reaction page

if __name__ == "__main__":

    app.run(debug=True,port=5001)

```

## Appendix C: Prediction

```
paths = []

labels = []

for dirname, _, filenames in os.walk('/kaggle/input'):

    for filename in filenames:

        paths.append(os.path.join(dirname, filename))

        label = filename.split('_')[-1]

        label = label.split('.')[0]

        labels.append(label.lower())

    if len(paths) == 2800:

        break

print('Dataset is Loaded')

## Create a dataframe

df = pd.DataFrame()

df['speech'] = paths

df['label'] = labels

df.head()

df = df[df['label'] != 'disgust']
```

```

df['label'].value_counts()

labels = np.array(df['label'].value_counts().index)

emotion = 'fear'

path = np.array(df['speech'][df['label']==emotion])[0]

data, sampling_rate = librosa.load(path)

#display the wave of the audio file

librosa.display.waveshow(data)

#display the spectrogram of audio file

spectrogram(data, sampling_rate, emotion)

Audio(path)


emotion = 'angry'

path = np.array(df['speech'][df['label']==emotion])[1]

data, sampling_rate = librosa.load(path)

librosa.display.waveshow(data)

spectrogram(data, sampling_rate, emotion)

Audio(path)


# Audio feature extraction method

# Mel-frequency cepstral coefficients (MFCC)

```

```

# It extract audio feature

def extract_mfcc(filename):

    y, sr = librosa.load(filename, duration=3, offset=0.5)

    mfcc = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=40).T, axis=0)

    return mfcc

# display the MFCC of 0th index audio

extract_mfcc(df['speech'][0])

X_mfcc = df['speech'].apply(lambda x: extract_mfcc(x))

X = [x for x in X_mfcc]

X = np.array(X)

X.shape

# 2800 audio files and 40 features of audio file

# convert the feature into 1 dimention

X = np.expand_dims(X, -1)

X.shape

# Mapped label into numbers example (sad -> 0 , happy -> 1)

from sklearn.preprocessing import OneHotEncoder

enc = OneHotEncoder()

```

```

y = enc.fit_transform(df[['label']])

y = y.toarray()

y.shape

# show the label mapped values

label_index_mapping = {i: label for i, label in enumerate(enc.categories_[0])}

print(label_index_mapping)

#training the model

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,
random_state=42)

# Train the model total 50 epochs

epoch = 50

history = model.fit(X_train, y_train, validation_split=0.2, epochs=epoch,
batch_size=64)

y_pred = model.predict(X_test)import numpy as np

y_pred_labels = np.argmax(y_pred, axis=1)

y_test_labels = np.argmax(y_test, axis=1)

from sklearn.metrics import accuracy_score

accuracy = accuracy_score(y_test_labels, y_pred_labels)

```



```
print(accuracy*100,"% accuracy")
```

```
pred_X_mfcc = pred_df['speech'].apply(lambda x: extract_mfcc(x))
```

```
pred_X = [x for x in pred_X_mfcc]
```

```
pred_X = np.array(pred_X)
```

```
pred_X = np.expand_dims(pred_X, -1)
```

```
model.predict(pred_X)[0]
```