Recognizing Emotion from Speech using Machine learning and Deep learning

By

Tonny Roy

ID: 191-15-12650

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised by

Dewan Mamun Raza

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

Co-Supervised by

Mr. Raja Tariqul Hasan Tusher

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh

January 13, 2025

APPROVAL

This Project titled "**Recognizing Emotion from Speech using Machine learning and Deep learning**," submitted by **Tonny Roy, ID: 191-15-12650** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **13-01-2025**.

BOARD OF EXAMINERS

Md Tamim Ahad

Dr. Md. Taimur Ahad
Associate Professor & Associate Head
Chairman, Department of CSE
FSIT Daffodil International University

13/01/25

Mohammad Monirul Islam
Assistant Professor, Department of CSE
FSIT Daffodil International
University

Jakaria 13/01/25

Md. Jakaria Zobair
Lecturer, Department of CSE,
FSIT Daffodil International
University

13/01/2025

Nazibur Rahman
Technical Lead - Database Administrator
Telenor - Grameen Phone Account

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dewan Mamun Raza**, **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

12.01.2025

Dewan Mamun Raza
Assistant Professor
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by:

Mr. Raja Tariqul Hasan Tusher
Assistant Professor
Department of Computer Science and EngineeringDaffodil International University

Submitted by:

Tonny Roy
13-01-2025

Tonny Roy
Student ID:191-15-12650
Department of Computer Science and EngineeringDaffodil International University

Department of Computer Science and EngineeringDaffodil University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Dewan Mamun Raza Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **"Machine Learning and deep learning"** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

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

ABSTRACT

In the analysis of psychological disorders, behavioral decision making, human machine interaction application speech recognition is plays a essential role. Speech emotion recognition is a system that detects emotions from live audio. people from all over the world utilize words to express their emotions, regardless of their origin. In this project, we focus on using machine learning (ML), which employs a dataset and algorithms to predict or detect any future possibilities. The data sets of audio files in wave format with 8 emotional states: anger, disgust, fear, happiness, pleasant, surprise, sadness, and neutral. Using the librosa library, features were extracted from the audio files in the datasets. The features were applied to multiple machine learning models and results were compared. Speech Emotion Recognition is a popular study topic with numerous applications. It has also became a challenge in the field of speech recognition processing too. Overall, a CNN model would be a good method to human speech emotion recognition with the accuracy rate 85%, because of its capacity to extract complicated patterns and characteristics from input data. The other two models accuracy rates are, SVM 82% and MLP 83%. However, the model's success would be determined by the quality of the preprocessed data, the model architecture used, and the efficacy of the data augmentation strategies employed

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The goal of identifying the underlying emotion in spoken language is known as Speech Emotion Recognition (SER). A vital component of human communication, the capacity to identify emotions in speech has numerous real-world uses in fields such a market research, education, virtual assistant analytics etc. Because emotions are highly contextual and subjective which is challenging. A speaker's emotional state can be change rapidly. Since emotions are highly contextual and subjective, SER is a challenging task. Different emotions can be expressed with the some words and speaker's emotional state can change quickly. As a result, SER calls for the application of advanced machine learning algorithms that are able to precisely identify the underlying emotions and capture the speech. A speech emotion dataset is suggested as a solution to this challenge in order to promote studies and applications in emotion analysis and recognition. This **RAVDESS** dataset includes voice recordings of 24 actors and IN TESS dataset, there are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant, surprise, sadness, and neutral). There are 2800 data points (audio files) in total, which are in English language. The production of this dataset aims to stimulate research and development in the field of emotion recognition and analysis by offering resource for training and evaluating Speech Emotion Recognition algorithms. Machine learning refers to a computer's ability to automate task using data and learning methods. Machine learning algorithms have been widely used for Speech Emotion Recognition. The support vector machine (SVM) is used as classifier to classify different emotional states as anger, happiness, sadness, neutral, fear from emotional database. We also use CNN and MLP classifier. Various methods have been effectively used for speech emotion recognition. In this study there were used a variety of machine learning methods, neural networks and sound wave characteristics for speech recognition of emotions.

1.2 Motivation

Speech Emotion Recognition has a wide range of potential uses, such as enhancing human-computer interaction by empowering computers to identify and react to emotions, enhancing customer service by analyzing emotions in call center conversations, supporting psychotherapy by assisting therapists in assessing their patient's emotional states, analyzing customer sentiments towards goods and services SER has many applications of various fields. These are- Virtual assistant, psychotherapy, education, market research and many more. The connection between consumers and virtual assistants like Google Assistant, Alexa can be improved using Speech Emotion Recognition system. Teachers can detect students who are in emotional distress and provide appropriate support by analyzing the emotional content of their speeches.

1.3 Objectives

The following are the main goals of speech emotion recognition:

- The basic goal of speech emotion recognition identification and classification is to recognize and classify the emotions that are communicated in spoken language. this includes feelings like surprise, fear, joy, sadness etc.
- Improving human computer interaction: By allowing computers to identify and react to human emotions, spoken emotion detection can enhance human-computer connection. A computerized customer support representative, for instance, can modify their response according to the customer's emotional state.
- motion recognition technology has the potential to improve speech-based therapies for people suffering from emotional disorders. For example, it can be used to monitor patients' emotional states during therapy sessions and alter treatment accordingly.
- Speech emotion recognition improves lie detection accuracy by identifying emotional indicators that suggest lying.

1.4 Methodology

This chapter will detail the research design, data collection methods, and analytical tops employed in the study. It will explain the rationale behind choosing machine learning and deep learning techniques, experimental setups, and data visualization process, providing a comprehensive understanding of the research methodology.

1.5 Project Outcome

Speech Emotion Recognition (SER) systems produce diverse and profound results, depending on the application domain. A main effect is the identification of emotional states such as happiness, sadness, anger, fear, or neutrality in spoken language, which allows for a better understanding of human communication. This can improve the user experience in virtual assistants, call centers, and customer service by making interactions more empathic and individualized. In healthcare, SER can help detect mental health concerns like sadness or anxiety early on by monitoring emotional indicators in speech. In education, it can enhance e-learning platforms by assessing students' emotional involvement. Furthermore, in entertainment and gaming, SER enhances interactive experiences by changing storylines or games to users' emotions.

Chapter 2

Background

2.1 Introduction

The process of identifying the emotional content of spoken language through computational methods is known as speech emotion recognition. Pitch, intensity, rhythm, timbre, and other acoustic characteristics of speech are analyzed, and machine learning algorithms are used to categorize the speaker's emotional state. There are several real-world uses for the capacity to identify human emotions from speech, such as boosting the precision of lie detection methods, creating more effective speech-based therapies for people with emotional problems, and increasing human-computer interaction. Typically, speech emotion identification systems use a huge dataset of labeled speech samples—where each speaker's emotional state is known—to train a machine learning model. New speakers' emotional states can then be instantly classified using the model. However, since emotions can be communicated in subtle and nuanced ways that are hard to capture using conventional acoustic features, effectively identifying human emotions from speech is a complex and difficult process. Therefore, the goal of current research in this area is to create increasingly complicated algorithms that can more accurately represent the intricacy and diversity of human emotional expression in speech.

2.2 Literature Review

The process of identifying the emotional content of spoken language through computational methods is known as speech emotion recognition. Pitch, intensity, rhythm, timbre, and other acoustic characteristics of speech are analyzed, and machine learning algorithms are used to categorize the speaker's emotional state. There are several real-world uses for the capacity to identify human emotions from speech, such as boosting the precision of lie detection methods, creating more effective speech-based therapies for people with emotional problems, and increasing human-computer interaction. Typically, speech emotion identification systems use a huge dataset of labeled speech samples—where each speaker's emotional state is known—to train a machine learning model. New speakers' emotional states can then be instantly classified using the model. However, since emotions can be communicated in subtle and nuanced ways

that are hard to capture using conventional acoustic features, effectively identifying human emotions from speech is a complex and difficult process. Therefore, the goal of current research in this area is to create increasingly complicated algorithms that can more accurately represent the intricacy and diversity of human emotional expression in speech.

The field of human speech emotion recognition has its roots in the study of human emotions and the development of computer-based systems for analyzing speech signals. In the early days of artificial intelligence and signal processing, researchers focused on developing rule-based systems that used expert knowledge and heuristics to classify emotions based on acoustic features of speech, such as pitch, intensity, duration, and spectral content.

In the 1990s and early 2000s, the field of speech emotion recognition began to shift towards machine learning and data-driven approaches, as researchers started using statistical models, such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), to classify emotions based on large datasets of speech signals and corresponding emotion labels.

In recent years, with the availability of deep learning algorithms and large-scale datasets, researchers have achieved significant progress in developing neural network-based models for speech emotion recognition.[7] These models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, can learn complex and abstract features from speech signals and achieve state-of-the-art performance on various emotion recognition tasks.

Today, human speech emotion recognition has applications in various domains, such as healthcare, education, entertainment, marketing, and security. It has the potential to revolutionize how humans interact with machines and enable more natural and intuitive communication interfaces. However, there are still challenges and limitations in the field, such as the lack of standardized datasets and evaluation metrics, the influence of cultural and contextual factors on emotion perception, and the ethical and privacy concerns related to the use of speech data for emotion recognition.

2.2.1 Similar Applications

The topic of Speech Emotion Recognition (SER) has seen a lot of research and applications in recent years, using a variety of approaches and technologies to increase accuracy and practicality. To categorize emotions based on speech characteristics including pitch, tone, energy, and speech rate, researchers frequently integrate machine learning methods like support vector machines (SVM), deep neural networks (DNN), and recurrent neural networks (RNN). While SER has been used in customer service to enhance call center performance by evaluating

caller mood, case studies in industries such as healthcare have investigated its use in identifying patients' emotional states. In terms of methodology, a sizable amount of research focuses on dataset generation, including openly accessible databases like TESS and RAVDESS, and feature extraction approaches (such as Mel-frequency cepstral coefficients and prosodic characteristics). Virtual assistants (like Apple's Siri and Google's Assistant) and mental health apps (like Woebot) are examples of web and mobile applications that use SER. These programs use users' emotional tones to tailor their responses. In order to better comprehend and address user emotions, these apps frequently combine SER with sentiment analysis and natural language processing (NLP). Overall, better emotion classification algorithms, a wider variety of emotion datasets, and integration with AI-driven applications in various industries are characteristics of SER developments.

2.2.2 Related Research

Speech Emotion Recognition (SER) has been extensively studied due to its applications in areas such as virtual assistants, education, and market research. Various approaches have been explored to address the complexities of recognizing emotions in speech, given their contextual and subjective nature. Joshi et al. (2013) analyzed different machine learning techniques for SER, emphasizing the challenges posed by rapidly changing emotional states and the need for robust datasets and algorithms (1). Lieskovská et al. (2021) highlighted the potential of deep learning and attention mechanisms in capturing intricate emotional patterns in speech. They demonstrated how neural networks enhance the accuracy and adaptability of SER systems (2). Abbaschian et al. (2021) reviewed the use of deep learning techniques and sound wave analysis for SER. They stressed the role of diverse datasets in improving the training and evaluation of these models (3). Similarly, Saini et al. (2022) investigated the application of Support Vector Machines (SVM) and Random Forest classifiers, showcasing their effectiveness in classifying emotions like anger, happiness, and sadness from emotional databases (4). These studies highlight the evolution of SER from traditional machine learning approaches to more advanced deep learning models, supported by comprehensive datasets and innovative methodologies.

2.3 Summary

In this chapter there were discussed about the reason behind the recognition of speech emotion. There are also some similar applications of this research and related research papers.

Chapter 3

Research Methodology

3.1 Methodology/Requirement Analysis & Design Specification:

The methodology for developing a human speech emotion recognition project can involve the following steps:

● **Problem Definition:** Define the problem statement and research questions. Identify the main objectives, stakeholders, and potential applications of the project.

● **Data Collection:** Collect speech data that covers a range of emotions and contexts. Ensure that the data is representative, diverse, and annotated with ground truth labels.

● **Preprocessing:** Preprocess the speech data by applying techniques such as noise reduction, feature extraction, and normalization.[8] This can involve using software tools such as Librosa Once the data is divided into attributes and labels, the final preprocessing step is to divide data into training and test sets. The model selection library of the Scikit-Learn library contains the train_test_split method that allows us to divide data into training and test sets.

.

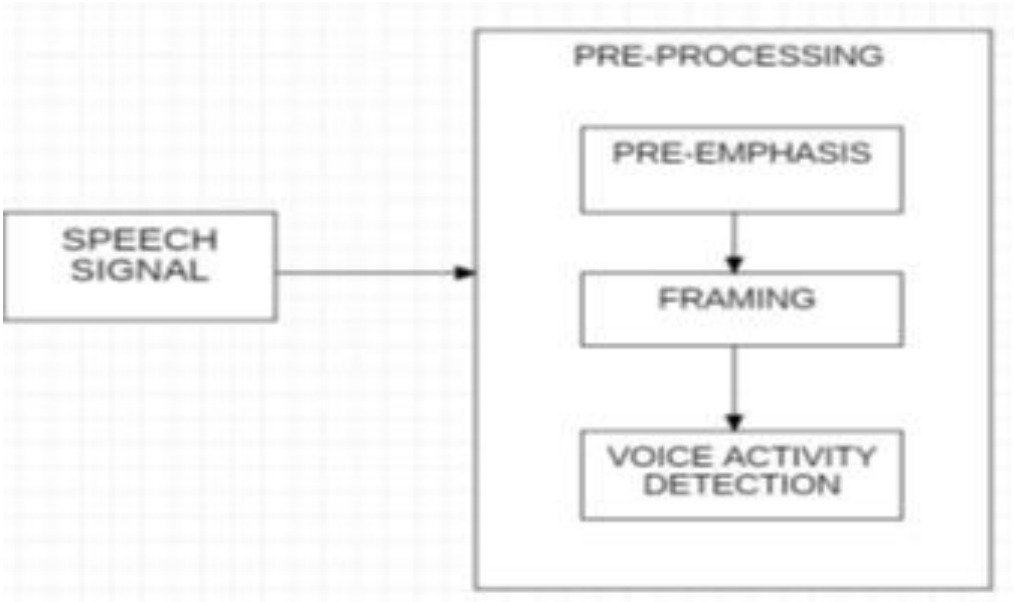

Figure3.1:data pre-processing steps

●**Feature Extraction :** The following speech features were extracted by chroma, cepstral coefficient (mfcc), and melspectogram (mel). These features are extracted and classified.

● **Model Selection:** Select an appropriate machine learning model or algorithm that can classify emotions based on speech learning.

● **Model Evaluation:** Evaluate the performance of the trained model using appropriate metrics such as accuracy, precision, recall, and F1-score. This can involve using techniques such as confusion matrices.

● **Model Optimization:** Optimize the model performance by fine-tuning the model architecture, regularization techniques, or feature selection methods.

● **Deployment:** Deploy the optimized model in a real-world setting. This can involve integrating the model with other technologies and systems, such as chatbots, virtual assistants, or gaming platforms.

● **Evaluation:** Evaluate the effectiveness and usability of the deployed model in a real-world setting. Collect feedback from end-users, stakeholders, and domain experts.

● **Maintenance:** Maintain the model by monitoring its performance, updating its parameters, and retraining it with new data. This can involve using techniques such as online learning, transfer learning, or active learning.

signals. This can involve using techniques such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN) and MLP(Multilayer perceptrons)

● **Model Training:** Train the selected model using the preprocessed speech data. This can involve using techniques such as cross-validation, hyperparameter tuning, or transfer

● **Label encoding:** Since you are using strings for your emotions, it is important that you use label encoding to transform these emotions into numbers; otherwise, your model just won't run!

3.1.1 Proposed Methodology/ System Design:

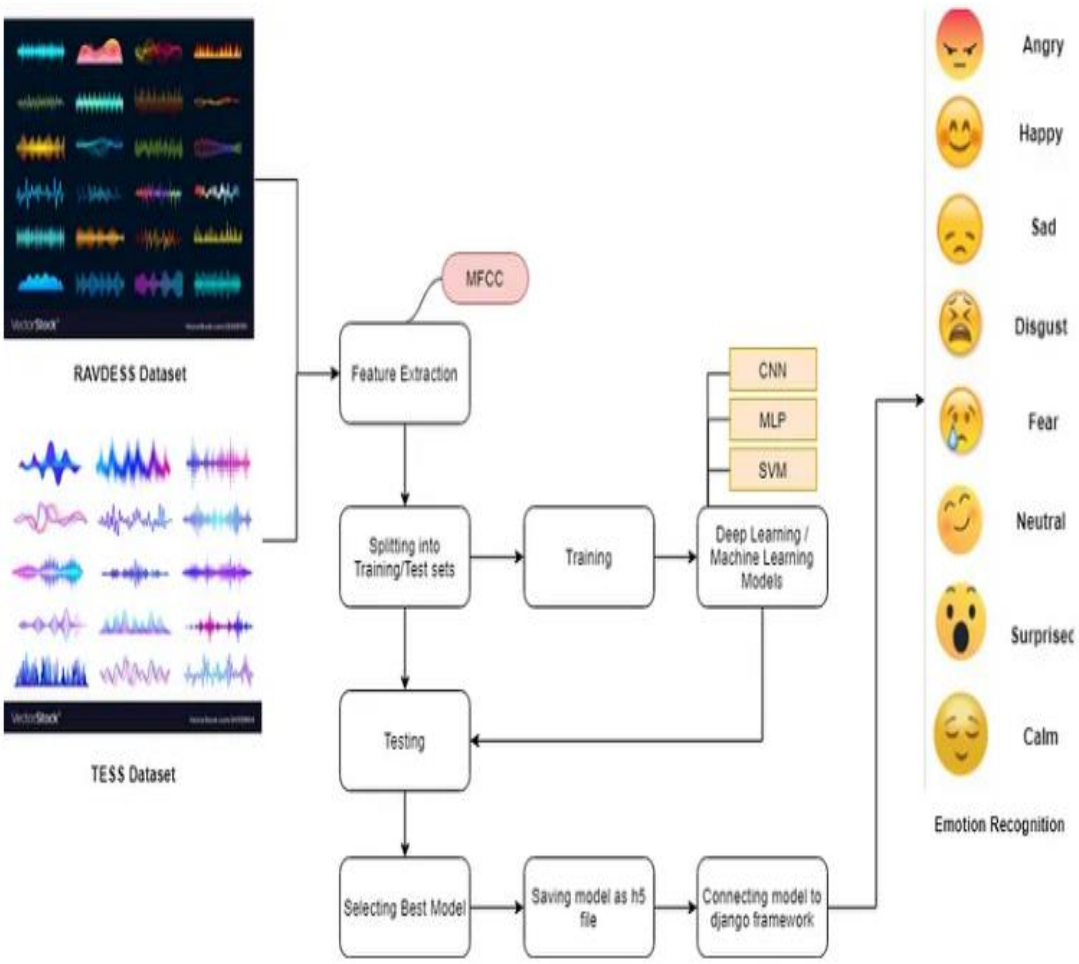

Figure 3.2: Proposed Overview SER system

3.1.2 Functional and Nonfunctional Requirements

Functional Requirements

These are the specific features and capabilities the system must provide.

1. Input Handling

The system must accept audio input in various formats (e.g., WAV, MP3, AAC).

The system should allow real-time audio input through microphones or pre-recorded files.

2. Preprocessing

Noise reduction and filtering should be performed on the audio input.

Support for normalization and resampling to ensure consistent input quality.

Feature extraction techniques like MFCCs, Chroma Features, and Spectral Features should be implemented.

3. Emotion Detection

The system must classify the input audio into predefined emotion categories (e.g., happy, sad, angry, neutral, etc.).

It should support multi-class classification for recognizing multiple emotions.

Optionally, the system could detect and handle mixed emotions or emotion intensity levels (e.g., low, medium, high intensity).

4. Output

Display the detected emotion(s) as text or visual indicators (e.g., graphs or labels).

5. Data Handling

Provide functionality to load, store, and manage audio datasets for training and testing.

Support data augmentation to improve the robustness of the model.

6. Model Training and Testing

Allow users to train custom models with labeled datasets.

Provide evaluation metrics (e.g., accuracy, precision, recall) after testing.

7. Multi-Language Support

If required, recognize emotions across different languages and accents.

8. Real-Time Processing

Enable real-time or near-real-time emotion recognition for live audio streams.

Non-Functional Requirements

These are the quality attributes and constraints of the system.

1. Performance

The system must process audio input and return results within a specified time frame

Maintain high accuracy (e.g., >85%) in emotion detection under ideal conditions.

2. Scalability

The system should handle an increasing volume of data and users without degradation in performance.

Support both local and cloud-based deployment for scalability.

3. Reliability

The system must perform consistently under various conditions, such as background noise or different accents.

Handle audio input gracefully when the quality is poor or the format is unsupported.

4. Usability

Provide an intuitive and user-friendly interface for non-technical users.

Include documentation and tutorials to guide users in using the system.

5. Security

Ensure the confidentiality of user data, especially if personal or sensitive audio recordings are involved.

Implement access control mechanisms to prevent unauthorized access.

6. Compatibility

Support integration with other systems and platforms through APIs or libraries.

Be compatible with popular operating systems like Windows, macOS, and Linux.

7. Maintainability

The system should have modular code that is easy to update or extend.

Use standard practices for software development, including clear documentation and version control.

8. Portability

The system should be deployable on multiple devices, such as desktop computers, servers, and mobile devices.

3.1.3 Work Flow Diagram Level

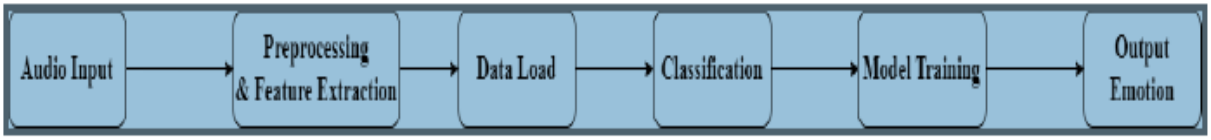

Figure 3.3: steps of work flow

●**Speech Input:** The initial stage is to upload and record the user's audio input.

● **Preprocessing:** Audio data is preprocessed to reduce background noise and extraneous sounds that could interfere with emotion recognition.

● **Feature Extraction:** Preprocessed audio data is evaluated to extract key aspects including pitch, loudness, and tone.

● **Classification:** Classification involves using extracted characteristics to train a machine learning or deep learning model on a labeled emotion dataset. The model assigns the auditory input to one or more.

● **Output Emotion:** The recognized emotion is outputted to the user or the application because of the classification step.

3.2 Detailed Methodology and Design

In the case of Speech Emotion Recognition, the 3.4 diagram can show the sub-processes involved in preprocessing, feature extraction, feature selection, classification, and post-processing. It also shows the inputs, outputs, and data flows between these subprocesses. Data Flow Diagrams are graphical representations of a system that show how data flows through different processes and entities. In the context of Human Speech Emotion Recognition, DFDs can be used to represent how the system processes and analyzes human speech to recognize the emotional state of the speaker.

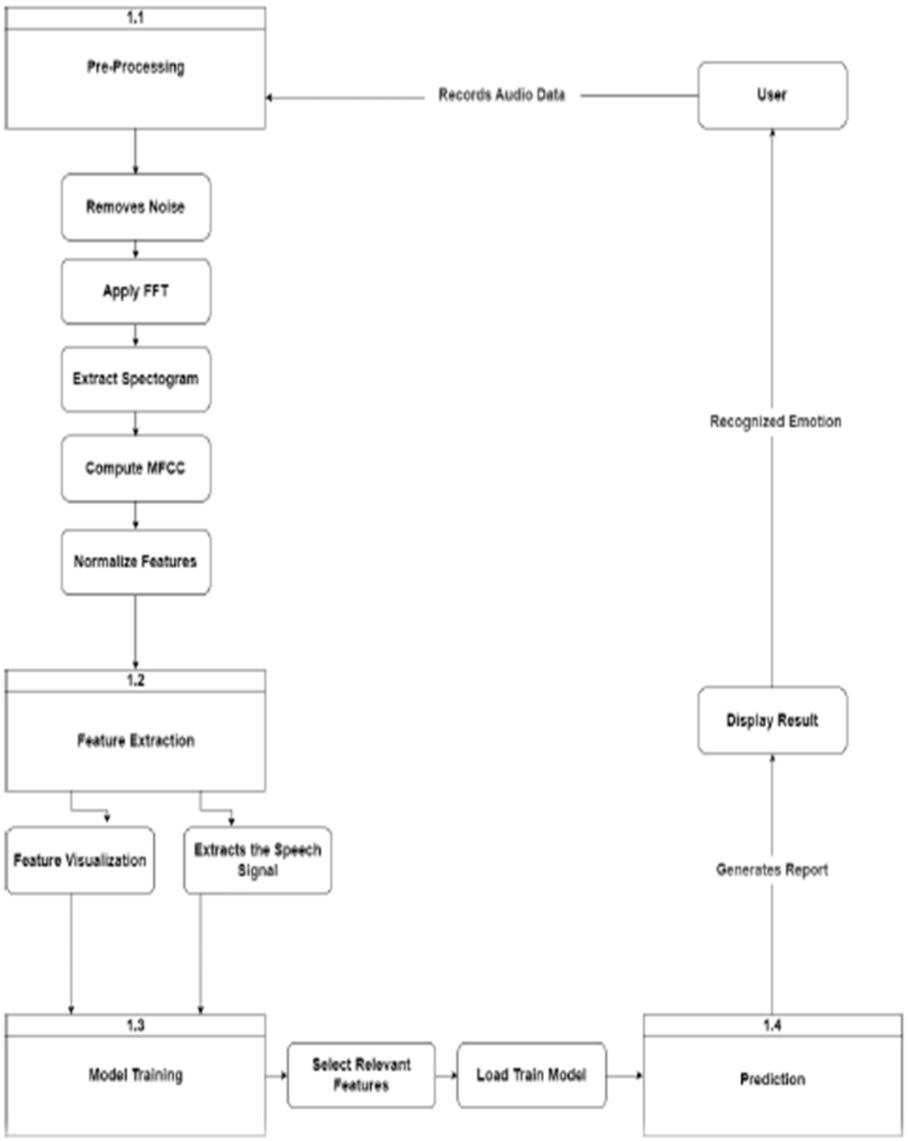

Figure 3.4: design of data flow

Chapter 4

Implementation and Results

4.1 Environment Setup

Setting up an environment for Speech Emotion Recognition (SER) often entails gathering the essential tools, libraries, and dependencies for processing audio input, extracting features, and training/analyzing models. The following is a step-by-step guidance for the environment setup:

1. Choose a Programming Language
2. Install Python
3. Install Required Libraries
4. Install Speech Emotion Recognition Datasets
5. Verify Installation
6. Set Up a Feature Extraction Workflow
   i. MFCC (Mel-Frequency Cepstral Coefficients): Commonly used for SER.
   ii. Chroma Features: Represent pitch-related information.
   iii. Spectral Features: Analyze spectral changes in audio.

7.Set Up Model Development

8.Test the Workflow

4.2 Testing and Evaluation Analysis

After completing pre processing and feature extraction of the data set I did try to build a Decision Tree Classifier to see how this type of machine learning model would perform.

The proposed model of speech emotion detection is based on a deep learning technique that includes convolutional neural networks(CNN), Support Vector Machine classifiers (SVM),and MLP classifiers. The important concept is to train the model using MFCC is known as “spectrum of a spectrum” .MFCC is a variant of the Mel-frequency cepstrum (MFC) that has been shown to be the state of the art in sound formalization in automatic speech recognition tasks. The MFC coefficients have been widely employed because to their ability to express the amplitude spectrum of a sound wave incompact vectorial form.

The audio file is separated into frames using a fixed window size, to obtain statistically stationary waves. The amplitude spectrum is normalized with a reduction of the "Mel" frequency scale. This operation is performed for empathizing the frequency more meaningful fir a significant reconstruction of the wave as the human auditory system can perceive.

For each audio file, 40 features were retrieved. The functionality was created by turning each audio file into a floating-point time series. The time series was then used to generate an MFCC sequence.

Decision Tree Classifier:

As part of the analysis, I did try to build a Decision Tree classifier to make a first attempt in accomplishing this classification task .

Surprisingly, the model got a 68% accuracy.

	precision	recall	f1-score	support
0	0.82	0.79	0.80	190
1	0.61	0.54	0.57	117
2	0.65	0.65	0.65	266
3	0.72	0.75	0.73	246
4	0.71	0.73	0.72	265
5	0.64	0.65	0.65	246
6	0.59	0.66	0.62	202
7	0.69	0.60	0.64	202
accuracy			0.68	1734
macro avg	0.68	0.67	0.67	1734
weighted avg	0.68	0.68	0.68	1734

Figure 4.1: Accuracy of Decision Tree Classifier**.**

CNN (convolutional neural networks):

The deep neural network (CNN) created for the classification challenge is reported operationally. For each audio file provided as a input, the network can work with 40 features vectors. The 40 numbers represent the concise numerical form pf a second audio frame . To run one round of a 1D CNN with ReLu activation function ,20%dropput, and a max pooling function 2*2, we use a data set of size<number of training files>40x1.

The rectified linear unit (ReLu) can be written as g(z) = max{0, z}, allowing us to achieve a big value in case of activation. This function is an excellent choice to represent hidden units. Pooling

can assist the model focus primarily on the key characteristics of each chunk of data, making them position-invariant. We repeated the technique outlined above, this time adjusting the kernel size. Following that, we applied another dropout and flattened the result to ensure compatibility with the subsequent layers.Finally, we used one Dense layer (fully connected layer) with a softmax activation function, varying the output size from 640 to 8, and estimating the probability distribution of each properly encoded class (0=Neutral, 1= Clam, 2= Happy, Sad=3, Angry=4, Fearful=5, Disgust=6, Surprised=7).

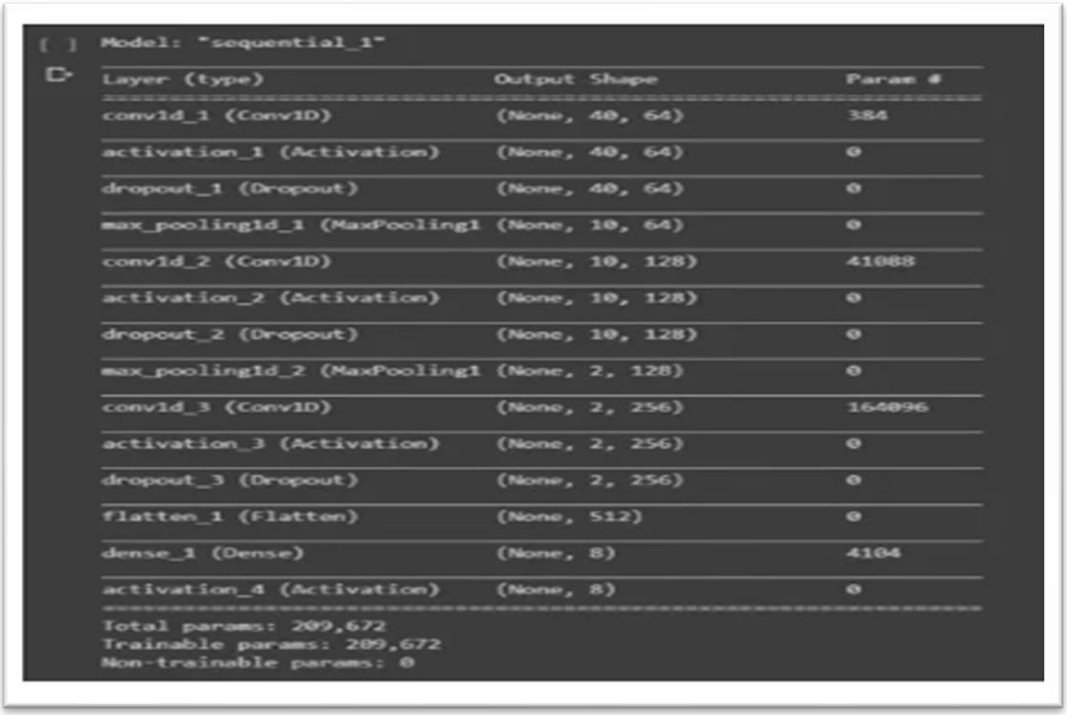

Figure 4.2: classification layer of CNN model

MLP(Multilayer perceptrons):

MLP classifier is an in-built classifier model that is included by default in the scikit learn library, which provides a large number of in-built classifiers. The result dataset from the feature extraction module is fed into the MLP classifier model. To perform classification, the dataset is first divided into test and train datasets. 10% of the original dataset is used for testing, while the remainder is used for training. The MLP classifier model is trained on both the test and training datasets and its accuracy is compared to each other. Since the MLP classifier is mostly recognized for hidden layersThe high number of hidden layers improves model correctness and performance.

SVM (Support Vector Machine):

"Support Vector Machine" (SVM) is a supervised machine learning technique that can be used for classification and regression tasks. However, it is mostly employed for categorization difficulties. The SVM algorithm plots each data item as a point in n-dimensional space (where n is the number of features), with the value of each feature representing the value of a certain coordinate.Data can be scaled before being fed into an SVM classifier in order to avoid processing attributes in larger numeric ranges. Scaling also helps to minimize numerical issues during calculations.

4.3 Results and Discussion

The success of this research is determined by the accuracy with which recognizes live audio. The execution depends on the dataset. We tried with different training and testing model ratios to achieve the best results. The ratio of 75% data for training and 25% for testing produced better results than other ratios. We employed a grid confusion matrix with MLP, CNN, support vector machine(SVM). The unweighted accuracy was derived as the average of every emotion accuracy.

CNN: In this step ,I build and train CNN model with our clean and prepared dataset .For this model , I decided to use 1D CNN as we have a time dimension aspect in our audio features. 1D CNN runs along one dimension and can take advantage of the audio wave's time structure.

```
[[172   5   1   8   0   0   1   3]
 [  6  89   5   9   0   0   6   2]
 [  4   6 223   5   8   8   3   9]
 [  7   7   3 212   3   8   4   2]
 [  1   2   4   2 232   8  15   1]
 [  0   1   6  23  10 196   8   2]
 [  1   2   0   2   5   1 186   5]
 [  5   4   6   4   2   1   8 172]]
```

Figure 4.3: Confusion matrix of CNN model

	precision	recall	f1-score	support
0	0.88	0.91	0.89	190
1	0.77	0.76	0.76	117
2	0.90	0.84	0.87	266
3	0.80	0.86	0.83	246
4	0.89	0.88	0.88	265
5	0.88	0.80	0.84	246
6	0.81	0.92	0.86	202
7	0.88	0.85	0.86	202
accuracy			0.85	1734
macro avg	0.85	0.85	0.85	1734
weighted avg	0.86	0.85	0.85	1734

Figure 4.4: Results of the CNN model on the test set per each class

SVM: Using the training data to train the SVM classifier. After that we was perform a classification task. Confusion matrix, precision, recall, and F1 measures are the most commonly used metrics for classification tasks. Scikit-Learn's metrics library contains the classification_report and confusion_matrix methods, which can be readily used to find out the values for these important metrics.

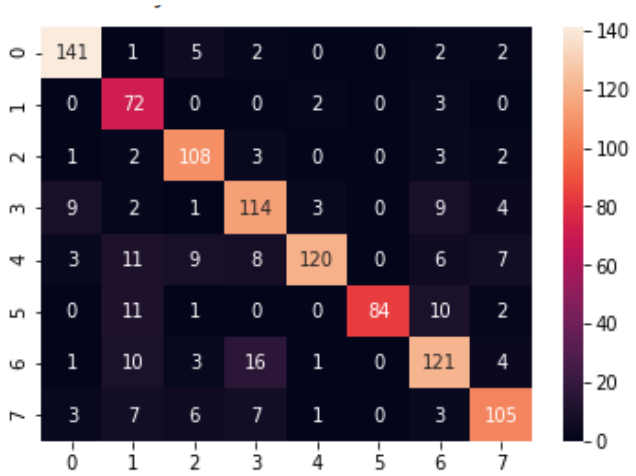

Figure4.5 : Confusion matrix of SVM Model

	precision	recall	f1-score	support
angry	0.89	0.92	0.91	153
calm	0.62	0.94	0.75	77
disgust	0.81	0.91	0.86	119
fear	0.76	0.80	0.78	142
happy	0.94	0.73	0.82	164
neutral	1.00	0.78	0.88	108
sad	0.77	0.78	0.77	156
surprised	0.83	0.80	0.81	132
accuracy			0.82	1051
macro avg	0.83	0.83	0.82	1051
weighted avg	0.84	0.82	0.82	1051

Figure4.6 : Results of the SVM model on the test set per each class

MLP : Multi-Layer Perceptrons is considered to the suitable model for speech emotion recognition.Because ,MLP classifier is the best suitable for complex datasets when compared to other models. After train the MLP classifier model we can predict the values for the test set. In the figure and figure here we can see the prediction and confusion matrix of the MLP classifier.

```
[[168   1   7  13   6   1   0   3]
 [  0  62   0   0   6  14   4   0]
 [  2   2 116   0   8   6   1   7]
 [  4   2   1 146   7   6   7   3]
 [  4   5   3   3 156  11   3   1]
 [  0   3   0   0   0 159   3   0]
 [  0  11   3   6   7  10 160   2]
 [  0   0   4   1  13  14   1 127]]
```

Figure4.7 : Confusion matrix of MLP classifier

	precision	recall	f1-score	support
angry	0.94	0.84	0.89	199
calm	0.72	0.72	0.72	86
disgust	0.87	0.82	0.84	142
fearful	0.86	0.83	0.85	176
happy	0.77	0.84	0.80	186
neutral	0.72	0.96	0.82	165
sad	0.89	0.80	0.85	199
surprised	0.89	0.79	0.84	160
accuracy			0.83	1313
macro avg	0.83	0.83	0.83	1313
weighted avg	0.84	0.83	0.83	1313

Figure4.8 : Results of the MLP model on the test set per each class

4.4 Summary

In this paper, we offer a deep neural network-based architecture for emotion categorization utilizing audio recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS). The model was trained to categorize seven different emotions (neutral, calm, happy, sad, furious, afraid, disgusted, startled) and received an overall F1 score of 0.85, with the best performance in the Happy class (0.90) and the weakest in the Calm class (0.77). To achieve this outcome, we extracted MFCC features (spectrum of a spectrum) from the audio files used in the training.

CLASS	MLP	SVM	CNN
SAD	0.81	0.82	0.80
ANGRY	0.89	0.91	0.89
HAPPY	0.82	0.84	0.90
DISGUST	0.80	0.81	0.81
SURPRISE	0.80	0.87	0.88
NEUTRAL	0.89	0.93	0.88
CALM	0.75	0.64	0.77
FEAR	0.84	0.81	0.88

Table 4.4.1: Evaluation Metrics of MLP, SVM and CNN model

MLP Classifier achieved an F1 score of 0.83 over the 8 classes.
SVM classifier achieved an F1 score of 0.82.
Final choice — deep learning model that obtained an F1 score of 0.85

Chapter 5

Engineering Standards and Design Challenges

5.1 Compliance with the Standards

Compliance with Speech Emotion Recognition (SER) standards assures that systems are dependable, secure, ethical, and perform well. Adherence entails complying with software quality frameworks such as ISO/IEC 25010 to ensure usability and reliability, implementing data annotation and management practices in accordance with FAIR principles and W3C Speech Interface Standards, and utilizing AI and machine learning guidelines such as ISO/IEC 23053 to promote transparency and reproducibility. Furthermore, compliance with audio processing protocols such as ITU-T P.56 and IEEE 269 provides high-quality input data, while ISO/IEC 27001 and GDPR protect sensitive user data. Finally, putting human-centered design principles into standards like ISO 9241 ensures usability, while thorough evaluation using methods like NIST SCTK certifies system accuracy and performance.

5.1.3 Software Standards

A range of methods and guidelines are included in the software standards for speech emotion recognition (SER) systems to guarantee their dependability, usefulness, and moral behavior. While ISO/IEC 12207 describes the lifecycle processes for software creation and maintenance, ISO/IEC 25010 deals with software quality attributes like performance efficiency, compatibility, and security. Additionally, SER systems need to adhere to data and AI standards, such as ISO/IEC 23053 for AI lifecycle frameworks and the FAIR principles for data management. Standards like ISO/IEC 14496-3 for audio encoding and ITU-T P.56 for voice level measurement are necessary to guarantee audio quality. Security and privacy, including adherence to frameworks like GDPR, HIPAA, and ISO/IEC 27001 for sensitive data, are essential.

5.1.2 Hardware Standards

The primary objective of hardware standards for Speech Emotion Recognition (SER) systems is to guarantee the reliability, effectiveness, and interoperability of technologies used to capture, process, and analyze speech data. High-quality microphones and audio

recording equipment must adhere to standards such as ITU-T P.51 for handset and headset testing or IEC 61672 for sound level meters in order to ensure accurate and noise-free audio input. IEEE 1241, which sets performance standards for ADCs, should be followed by ADCs and other signal processing equipment.

5.1.3 Communication Standards

The goal of communication standards for speech emotion recognition (SER) is to guarantee that the technologies and systems used to detect emotions in speech are morally sound, efficient, and interoperable with other communication-based applications. By preserving accuracy, security, and privacy, these standards seek to maximize human-machine interaction. These consist of data exchange methods, privacy laws, and moral considerations. Standardized data formats, voice signal processing methods, and cross-platform interoperability are important components.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Societal Impact

Speech Emotion Recognition (SER) research has a huge social impact since it improves human-computer interaction and promotes technological inclusion. By allowing robots to recognize and respond to human emotions, SER promotes more empathic and adaptive systems in areas such as mental health, education, and customer service. For example, it can help identify early indicators of emotional distress and provide timely care to people suffering from mental health concerns. In education, SER can personalize learning sessions by identifying students' emotional states, hence increasing engagement and outcomes. Furthermore, the technology improves accessibility for those with disabilities by providing more intuitive communication interfaces. Regardless of its promise, ethical concerns about privacy and data usage must be addressed to enable responsible deployment and prevent abuse, such as emotional profiling or monitoring.

5.2.2 Environmental Impact

While our project on data breach detection and prevention using machine learning and encryption may not have a direct environmental impact, we recognize the potential for indirect contributions to environmental sustainability.

5.2.3 Ethical Aspects

Privacy: Emotion analysis is highly personal. Unauthorized data acquisition and analysis can violate privacy.

Misuse: SER may be used for monitoring , manipulation, or unfair assessments.

Transparency: Lack of transparency in complex models might make it more difficult to detect bias and comprehend them.

Societal Impact: widespread use could impact autonomy, social interactions, and mental health.

5.2.4 Sustainability Plan

Speech Emotion Recognition sustainability is dependent on reducing computational energy usage, optimizing algorithms and using renewable energy infrastructure. By focusing on energy-efficient models and sustainable hardware methods , SER research can reduce its environmental impact while retaining technological advancement.

5.3 Project Management and Financial Analysis

The budget enables a business to use open-source tools, cloud-based infrastructure, and a smaller workforce to drastically cut initial costs. With this strategy, entrepreneurs or smaller businesses with little funding can now more easily acquire SER technology. The company can achieve operational efficiency without making a significant upfront investment by concentrating on modular development and implementing scalable infrastructure. Additionally, the ability to scale gradually without sacrificing system quality is made possible by utilizing pre-existing datasets and reducing data collecting expenses.

5.4 Complex Engineering Problem

Speech Emotion Recognition (SER) is a challenging field with numerous complex engineering problems. Here are some of the most prominent ones:

1. Subjectivity and Variability of Emotions:

Cross-cultural Differences: Emotions are expressed differently across cultures, making it difficult to generalize models trained on one culture to another.

Intra- and Inter-speaker Variability: The same emotion can be expressed differently by different individuals (inter-speaker variability) and even by the same individual at different times (intra-speaker variability).

Contextual Dependence: Emotions are often context-dependent, meaning the same utterance can convey different emotions depending on the situation.

2. Data Challenges:

Limited and Imbalanced Datasets: Most publicly available datasets are relatively small and often suffer from class imbalance, where some emotions are underrepresented.

Data Quality: Noise, accents, and varying recording conditions can significantly impact the

accuracy of SER systems.

Data Annotation: Accurately labeling emotions in speech data is subjective and time-consuming, requiring expert human annotators.

3. Feature Extraction and Selection:

Robust Feature Engineering: Identifying the most informative features from speech signals is crucial for accurate emotion recognition. This requires careful consideration of acoustic, prosodic, and linguistic features.

Dimensionality Reduction: High-dimensional feature spaces can lead to overfitting and computational inefficiency. Effective dimensionality reduction techniques are needed to select the most relevant features.

4. Model Development and Evaluation:

Model Complexity: SER models often involve complex deep learning architectures, which can be computationally expensive to train and deploy.

Overfitting and Generalization: Overfitting is a common problem in SER, especially when dealing with limited data. Generalization to unseen data remains a challenge.

Evaluation Metrics: Choosing appropriate evaluation metrics for SER is crucial, as traditional accuracy metrics may not fully capture the nuances of emotion recognition.

5. Real-world Applications:

Robustness in Real-world Environments: SER systems must be robust to noise, background interference, and other real-world challenges to be effective in practical applications.

Ethical Considerations: Privacy concerns and potential biases in data and models are important ethical considerations that must be addressed in SER research and development.

User Acceptance: Designing user-friendly and intuitive interfaces for SER systems is crucial for their successful adoption in real-world scenarios.

5.5 Complex Problem Solving

Speech Emotion Recognition (SER) presents numerous complex challenges that require innovative problem-solving approaches. Here's a breakdown of key issues and potential solutions:

1. Subjectivity and Variability of Emotions solution

Integrating visual cues (facial expressions, body language) with acoustic features can provide richer context and improve accuracy. Training models on diverse datasets and using techniques like domain adaptation to improve performance across different cultures.

2. Data Challenges solution

Techniques like noise injection, speed perturbation, and pitch shifting can artificially increase

dataset size and diversity. Prioritize labeling of the most informative samples to improve annotation efficiency. Pre-train models on large unlabeled datasets to learn robust representations.

3. Feature Extraction and Selection solution

Deep neural networks, especially convolutional and recurrent neural networks, can automatically learn hierarchical representations from raw audio data. Allow models to focus on the most important parts of the speech signal for emotion recognition.

4. Model Development and Evaluation solution

Dropout, early stopping, and weight decay can help prevent overfitting. Combining predictions from multiple models can improve robustness and accuracy. Consider metrics like confusion matrices, F1-scores, and receiver operating characteristic (ROC) curves to evaluate performance comprehensively.

5. Real-world Applications solution

Noise-Robust Feature Extraction: Techniques like spectral subtraction and wavelet transforms can mitigate the effects of noise. Conduct user studies to evaluate the effectiveness and usability of SER systems in real-world scenarios.

5.6 Engineering Activities

Speech Emotion Recognition (SER) is an interdisciplinary field that involves various engineering activities. Here are some key ones:

1. Signal Processing

Preprocessing**,** Segmentation, Normalization**,** Feature Extraction-Extracting relevant features from the audio signal that carry emotional information. Common features include:

Pitch, intensity, formants, Mel-Frequency Cepstral Coefficients (MFCCs), Duration, pauses, rhythm, intonation, etc.

2. Machine Learning

Model Selection: Choosing appropriate machine learning algorithms for classification. Common choices include:

Support Vector Machines (SVMs),Deep Neural Networks (DNNs), such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

Model Training: Training the chosen model on a labeled dataset of speech samples.

Model Evaluation: Assessing the performance of the trained model using appropriate metrics, such as accuracy, precision, recall, F1-score, and confusion matrices.

3. Software Engineering

System Design: Designing and developing SER systems that are efficient, scalable, and user-

friendly.

Implementation: Implementing the system using programming languages like Python, MATLAB, or C++.

Deployment: Deploying the system on various platforms, such as mobile devices, cloud servers, or embedded systems.

4. Data Engineering

Data Collection: Gathering and organizing large datasets of speech samples with accurate emotion labels.

Data Annotation: Manually labeling speech data with emotion categories.

Data Management: Storing, managing, and accessing large amounts of speech data efficiently.

5. Human-Computer Interaction (HCI)

User Interface Design: Designing intuitive and user-friendly interfaces for interacting with SER systems.

Usability Testing: Evaluating the usability and effectiveness of SER systems in real-world scenarios.

Ethical Considerations: Addressing ethical concerns related to privacy, bias, and fairness in SER systems.

5.7 Summary

Data quality, system reliability, and adherence to ethical and privacy standards are the primary objectives of Speech Emotion Recognition (SER) engineering standards. These standards provide guidelines for uniformity and excellence in fields like software development techniques, data processing, and the transparency of machine learning models. The complexity of accurately interpreting emotions from vocal cues alone, ambient noise, and speech variability caused by human traits (e.g., age, gender, accent) are some of the design challenges that SER faces. Additionally, developing culturally sensitive SER systems and ensuring their ethical use—particularly with regard to privacy and data protection—present a considerable challenge. Making sure SER systems perform well in a range of real-world situations while maintaining user trust and system transparency is another significant problem

Chapter 6
Conclusion

6.1 Summary

In conclusion, our approach utilizing audio recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS). The model was trained to categorize seven different emotions (neutral, calm, happy, sad, furious, afraid, disgusted, startled) and observed that CNN has highest accuracy with 85%, with the best performance in the Happy class (90%) and the weakest in the Calm class (77%). To achieve this outcome. Hence, CNN is the most efficient of the three models and provides the highest level of emotion analysis accuracy. Future research in Speech Emotion Recognition may focus on improving the accuracy and robustness of the system, exploring the use of multimodal data sources, and investigating the impact of cultural and linguistic differences on emotion recognition.

6.2 Limitation

- The system doesn't have multilingual support
- SER doesn't have any real time emotion detection feature

6.3 Future Work

- Extend the system's capabilities to recognize emotions in multiple languages
- Implement a real-time recording option for users to detect their own feeling.
- Record my own audio files and see how the model performs on other datasets.

References

[1] R. K. Aastha Joshi1, "A Study of Speech Emotion Recognition Methods," *International Journal of Computer Science and Mobile Computing* , pp. 28-31, 2013.

[2] M. J. J. a. C. Lieskovská *ORCID, "A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism," *Faculty of Electrical Engineering and Information Technology, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovakia*, May 2021.

[3] D. S.-S. a. A. E. Babak Joze Abbaschian, "Deep Learning Techniques for Speech Emotion Recognition,," *Computer Science and Engineering Department, University of Louisville, Louisville, KY 40292, USA;*, 2021.

[4] A. R. K. S. K. S. S. J. J. S. K. Anu Saini1, "An investigation of machine learning techniques in speech emotion recognition," *Indonesian Journal of Electrical Engineering and Computer Science*, pp. 252-260, Jun 26, 2022.

[5] K. S. K. T. P. A. K. L. B. B. I. &. G. S. V. S. K. (. Rao, "Speech Emotion Recognition with deep learning," pp. 252-260, 2020.

[6] C. A. D. S. S. Mr. Abhishek Kumar Saw, "International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.," *International Journal of Health Sciences,*, Vols. (S1),, pp. 14314-14321, 2022.

[7] P. D. D. S. Prof. Kinjal S. Raja1*, "Speech Emotion Recognition Using Machine Learning," *Educational Administration: Theory and Practice*, pp. 119-124, 2024.

Recognizing emotion from Speech using Machine learning and Deep learning

ORIGINALITY REPORT

23%	18%	15%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	medium.com Internet Source	9%
2	stackabuse.com Internet Source	2%
3	www.researchgate.net Internet Source	1%
4	Submitted to Higher Education Commission Pakistan Student Paper	1%
5	Submitted to PSG Institute of Management Coimbatore Student Paper	1%
6	Marco Giuseppe de Pinto, Marco Polignano, Pasquale Lops, Giovanni Semeraro. "Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients", 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2020 Publication	1%