BSC_Research_Project_Report final.pdf

WORD COUNT	5923	TIME SUBMITTED	19-MAY-2024 10:58AM
		PAPER ID	109058424

EmoSpeech: Identifying Emotions Through Speech Analysis Using MLP Classifier

In partial fulfilment of B.Sc. in Computer Science and Engineering with the degree requirements of this Project Report is submitted to the Department of Computer Science and Engineering, Jahangirnagar University

Submitted By

Md Omar Faruk

Exam Roll: 191345

Registration No: 20190648624

Arafat Alom

Exam Roll: 180686

Registration No: 46330

Session: 2018 - 2019

Supervised by

Sabrina Sharmin

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JAHANGIRNAGAR UNIVERSITY

May 2024

Abstract

Emotion detection in speech is a crucial aspect of human-computer interaction, with applications ranging from customer service chatbots to mental health monitoring systems. In this study, we explore the efficacy of using a Multilayer Perceptron (MLP) classifier to recognize emotions in speech data. Our dataset comprises 2800 audio samples sourced from the Toronto Emotional Speech Set (TESS)(EU JIN LOK, owner)[0], covering five primary emotions: 'happy,' 'sad,' 'disgust,' 'angry,' and 'fear.' We conducted a comprehensive analysis, including data preprocessing, feature extraction, model training, and evaluation.

To assess the performance of our model, we employed k-fold cross-validation, achieving an average accuracy of 77.62%. The dataset was partitioned into 80% training and 20% testing subsets to train and validate the MLP classifier. The precision values for the emotions 'disgust,' 'angry,' 'fear,' 'sad,' and 'happy' were found to be 0.947438, 0.953072, 0.947338, 0.947438, and 0.949371, respectively, while the corresponding recall values were 0.897500, 0.905000, 0.897500, 0.897500, and 0.902500. These results demonstrate promising capabilities in recognizing emotions from speech, with implications for various real-world applications. Future work could explore additional feature engineering techniques and alternative classification algorithms to further enhance the model's accuracy and robustness.

Declaration

The research work entitled **"EmoSpeech: Identifying Emotions Through Speech Analysis"** has been carried out in the Department of Computer Science and Engineering, Jahangirnagar University is original and conforms the regulations of this University.

We understand the University's policy on plagiarism and state that no part of this project has been copied from other sources or already submitted elsewhere for the award of any degree or diploma.

(Candidate 1)

(Candidate 2)

Counter Signed by

(Supervisor)

Acknowledgement

We would like to express our deepest thanks and gratitude to every person who has given the project their full attention and time. We owe a debt of gratitude to our supervisor, Sabrina Sharmin, for her direction, encouragement, and suggestions, all of which helped us tremendously in our effort to accomplish the year's culminating project. We admire our parents and the faculty of the Department of Computer Science and Engineering for encouraging us to pursue a career in computer science.

TABLE OF CONTENTS

LIST OF FIGURES

TERMINOLOGIES

AI *- Artificial Intelligence*
FN *- False Negatives*
FP *- False Positive*
ML *-Machine Learning*
TP *- True Positives*
TN *- True Negative*

Chapter 1
Introduction

This chapter contains some key information about the topic of our project. We have discussed the motivations for conducting this study, especially the background section: why we have chosen this topic, what is the problem here, what is the purpose of our research, and how we will contribute to our project.

1.1 Background Research

1.1.1 What is Speech emotion?

Emotion is the cornerstone of human interactions. In fact, the manner in which something is said can convey just as much information as the words being spoken. Emotional cues in speech are conveyed through vocal inflections known as prosody. Key attributes of prosody include the relative pitch, duration, and intensity of the speech signal. Together, these features encode stress, intonation, and rhythm, all of which impact emotion perception. While we have identified general patterns to relate prosody to emotion, machine classification and synthesis of emotional speech remain unreliable(Johns Hopkins University)[1].

Speech emotion recognition (SER) has traditionally relied solely on acoustic data for analysis. Acoustic features, extracted per frame, are typically processed using classifiers like support vector machines or multi-layer perceptrons in machine learning or deep learning frameworks, respectively, to assign emotion labels. However, research indicates that SER based solely on acoustic features often yields subpar performance. Alternatively, in addition to acoustic signals, linguistic information can also be derived from speech. Linguistic features are extracted from transcribed text using automatic speech recognition systems. Combining acoustic and linguistic information has the potential to enhance SER performance. This study conducts a comprehensive review of bimodal emotion recognition techniques that integrate both acoustic and linguistic data. It covers five key aspects of bimodal SER: emotion models, datasets, features, classifiers, and fusion methods. Notable findings, including state-of-the-art results and methodologies employed across commonly used datasets, are summarized to provide insights for ongoing research endeavors and to set benchmarks for future advancements. Furthermore, the survey identifies outstanding challenges in bimodal SER research, suggesting potential directions for future investigations.(Mladenovic, Nenad)[2]

Speech emotion recognition (SER) involves using artificial intelligence (AI) techniques to predict human emotions from audio signals. This technology finds applications across diverse fields including psychology, medicine, education, and entertainment. A critical aspect of SER is extracting relevant features from audio signals to accurately discern emotions(Oumaima Saidani, Rashid Jahangir)[3].

1.1.2 What is MLP?

An artificial neural network (ANN) is a computational model inspired by the structure and operation of the human brain's intricate network of interconnected neurons. It comprises interconnected units known as artificial neurons, which are organized into layers. Information propagates through the network, with each neuron processing input data and generating an output that influences other neurons in the network.

A multi-layer perceptron (MLP) is a specific type of artificial neural network that comprises multiple layers of neurons. These neurons commonly utilize nonlinear activation functions, enabling the network to discern intricate patterns within data. MLPs are highly valued in machine learning due to their capacity to comprehend nonlinear relationships in data, making them formidable models for tasks such as classification, regression, and pattern recognition. This tutorial aims to delve into the fundamental principles of MLPs and elucidate their internal mechanisms.(datacamp)[4].

Multi-layer perceptron, commonly abbreviated as MLP, comprises fully connected dense layers capable of transforming input dimensions into desired output dimensions. It is a type of neural network featuring multiple layers, where neurons are interconnected such that the outputs of some neurons serve as inputs to others. Each MLP typically includes an input layer with one neuron per input, an output layer with a single node for each output, and an arbitrary number of hidden layers, each containing varying numbers of nodes. The structure of a Multi-Layer Perceptron (MLP) is illustrated in the schematic diagram below.(geeksforgeeks)[5].

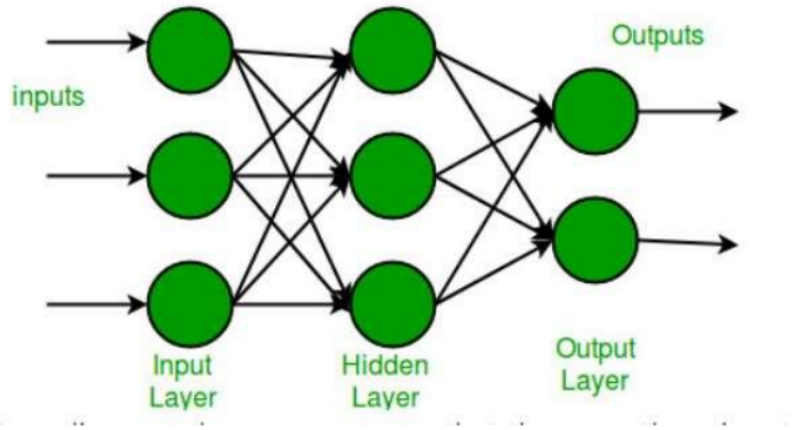


Figure 1.1: Layers of MLP classifier

1.2 Research Problem

For the last ten years, there has been a lot of work in speech emotion recognition. A lot of supervised and unsupervised machine learning has been applied. But because of the lack of a dataset, there have been many obstacles, like models giving less accuracy. Some of these works didn't show cross-validation, assessment of recall and precision. To improve it, researchers are applying deep learning techniques and neural networks like convolutional neural networks and recurrent neural networks. Although there has been work regarding speech emotion detection and it may have great accuracy, unless others, apart from researchers, can understand the inner factors that are affecting the model, this is not useful. Machine learning and deep learning algorithms are commonly used to identify speech emotion detection. however, because these models are opaque and difficult to interpret, it may be difficult to understand how they derive their predictions. Here we have developed a model using the MLP classifier though it was popular, we have amended this model from previous work.

Researchers utilize various audio processing methodologies to uncover this underlying layer of information, aiming to enhance and extract tonal and acoustic characteristics from speech.

The conversion of audio signals into numeric or vector formats presents complexities distinct from those encountered with images. The choice of transformation technique significantly impacts the preservation of crucial information when transitioning from the original "audio" format. Inadequacies in capturing subtleties like softness and tranquility can pose challenges for models tasked with learning emotions and classifying samples effectively.

Among the methods for converting audio data into numerical representations are Mel Spectrograms, which visualize audio signals according to their frequency components, allowing for plotting as an audio wave and utilization in training Convolutional Neural Networks (CNNs) as image classifiers. Another approach involves the extraction of Mel-frequency cepstral coefficients (MFCCs). Each of these data formats possesses its own advantages and drawbacks, contingent upon the intended application [6].

1.3 Research Objective

The objectives of this project encompass a multifaceted approach to advancing speech emotion recognition. Primarily, the aim is to develop a sophisticated system capable of accurately identifying and categorizing emotions conveyed through speech. This endeavor involves delving into diverse machine learning algorithms and techniques to refine the system's accuracy and reliability. Leveraging datasets like the TESS dataset, the project aims to train and validate the emotion recognition system comprehensively. Moreover, the project seeks to analyze the efficacy of different feature extraction methods and their impact on emotion recognition performance. By implementing cross-validation techniques, the generalization ability of the system can be thoroughly assessed. Precise evaluation metrics such as precision, recall, and overall performance are pivotal in gauging the system's effectiveness across various emotional states. Additionally, efforts are directed towards enhancing the interpretability of the system to discern underlying factors influencing emotion classification. Beyond technical exploration, the project aims to explore practical applications of the emotion recognition system across domains like human-computer interaction and sentiment analysis. The objectives of this project may include:

1. Developing a speech emotion recognition system capable of accurately identifying and classifying emotions expressed in speech.

2. Exploring various machine learning algorithms and techniques to improve the accuracy and reliability of emotion recognition.

3. Utilizing a diverse dataset, such as the TESS dataset, to train and test the emotion recognition system.

4. Investigating the impact of different features and feature extraction methods on emotion recognition performance.

5. Implementing cross-validation techniques to assess the generalization ability of the emotion recognition model.

6. Evaluating the precision, recall, and overall performance of the emotion recognition system for different emotions.

7. Enhancing the interpretability of the emotion recognition system to understand the underlying factors contributing to emotion classification.

8. Exploring potential applications of the emotion recognition system in various domains, such as human-computer interaction, healthcare, and sentiment analysis.

9. Providing insights into the limitations and challenges of current emotion recognition technology and proposing avenues for future research and development.

10. Ultimately, aiming to develop a robust and efficient speech emotion recognition system that can contribute to advancements in understanding human emotions and improving human-machine interaction.

1.4 Research Contributions

In this work, we have computed the accuracy of emotion detection based on TESS dataset and proposed a supervised learning model. We have preprocessed data and extracted 180 features, the main contributions of the paper are summarized as follows:

- We computed precsion for Angry, Sad, Happy, Disgust and Fear emotions.
- We computed Recall for Angry, Sad, Happy, Disgust and Fear emotions .We have performed K-Fold cross validation where k=5. Note that we have splited data into 5 slot where 4 slot were used for training and 1 was for testing.

- We have calculated average accuracy of our model based on testing data. Note that for five emotions we took 2000 data where 1600 data for training and 400 data for testing.

1.5 Research Applications

Simple speech recognition finds extensive use in various applications such as YouTube's auto-generated subtitles, live speech transcripts, online course transcripts, and voice-activated chatbots like Alexa and Siri. Research efforts in this area have led to significant improvements, evident in the yearly enhancements of YouTube's auto-generated subtitles. Conversely, speech emotion recognition (SER) applications introduce a new dimension to AI utilization, promising to further enhance our daily lives. One notable recent application of SER arises from the surge in online learning, enabling educators to gauge student responses in real-time and offer tailored assistance accordingly. Another emerging application involves assessing candidates for managerial roles by analyzing their emotional cues during audio or video interviews, providing hiring managers with valuable insights into candidates' suitability. Additionally, SER holds promise in evaluating the performance of existing employees, particularly in industries like call centers where interactions with customers are crucial. Implementing SER systems can enhance customer service efficiency, address grievances more effectively, and even monitor and support employee emotional well-being [7].

Chapter 2
Literature Review

For a long time, researchers in domains such as data mining, information retrieval, and natural language processing (NLP) have been focused on the automatic detection speech emotion. There has been a spike in interest in this sector as a result of the growth of social media and social platforms. In this part, we will look at the latest studies on detecting speech emotion in audio content using machine learning.

2.1 Speech Related Works

At UAIC-FII, two categories of courses are taught in different areas of natural language and speech processing. At the undergraduate level, the terminal year students can take an elective course on natural language processing (which presents mainly theories and techniques for discourse level representation and processing of language), and at the master level, the students in Computational Linguistics, over the two years program, take a general introductory course on computational linguistics, a theoretical course on syntax, a course on corpus linguistics, one in lexical semantics, one on machine translation and one on speech processing. The laboratory activities of both undergraduate and graduate courses make heavy use of corpora and other linguistic resources. In these activities students usually work in teams to perform term projects and homework. The goal of all projects is three fold: a). to train students to exploit existing corpora through the program interfaces available on the web, by integrating function calls in their own applications; b). to train students to make use of annotations applied to texts, by first manually annotating a small corpus and then devising their own tools that exploit the annotation for NLP applications; c). to build small annotated corpora, including spoken language (speech) corpora, and appropriate exploitation software that remain in the Faculty for further NLP research(Dan Cristea,2004)[8]

The documented Speech Recognition System relies on the CMUsphinx as its foundational API for acquiring speech recognition outcomes and is developed using Java. Its core objective is to empower users to tailor the speech recognition process according to their preferences. This customization encompasses defining models for the recognizer, specifying the processing of speech results, and

determining subsequent actions to be executed. Users can furnish these specifications through plugins, which are encapsulated classes conforming to provided interfaces and bundled into a jar file. The configuration file must include details of the classes designated for loading as modules. Leveraging these interfaces, users can effortlessly implement a wide array of functions using the provided plugin system.

The authors of this research proposed a technique that adjusts the classification outcomes using the stack classifier model in accordance with the compatibility of the section. The interior is said to be challenging to maneuver for frequent gatherings. In the end, this approach offers the desired classification outcomes simply and inexpensively(Mohammed Flaeel Ahmed Shariff,2015).[9]

Detecting emotions from speech presents a challenging task due to the nuanced nature of defining emotions. In this study, we employ a feature-engineering approach to address speech emotion recognition. By framing the issue as a multi-class classification problem, we assess the effectiveness of two model categories. In both approaches, we extract eight manually designed features from the audio signal. The first method involves utilizing these features to train six traditional machine learning classifiers, while the second approach employs deep learning, training a baseline feed-forward neural network and an LSTM-based classifier using the same features. To enhance communication clarity, we integrate features from the textual domain. Our evaluation includes metrics such as accuracy, f-score, precision, and recall across various experimental setups. Overall, our findings demonstrate that simpler machine learning models trained on a limited set of handcrafted features achieve performance comparable to current state-of-the-art deep learning methods for emotion recognition.(Gaurav Sahu,2019)[10].

Recognizing emotions from speech poses a significant challenge, and prevailing approaches heavily rely on models utilizing audio features to build effective classifiers. In our research, we introduce an innovative deep dual recurrent encoder model that simultaneously leverages text data and audio signals to enhance the understanding of speech data. Given that emotional dialogue comprises both auditory and linguistic elements, our model employs dual recurrent neural networks (RNNs) to encode information from audio and text sequences independently, then integrates these sources to predict emotion classes. This architecture allows for a comprehensive analysis of speech data, ranging from signal-level to language-level information, surpassing models focusing solely on audio features. Through extensive experiments, we examine the effectiveness and characteristics of our proposed model. When applied to the IEMOCAP dataset, our model demonstrates superior performance compared to previous state-of-the-art methods, achieving accuracies between 68.8% and 71.8% in

categorizing data into four emotion classes (angry, happy, sad, and neutral).(Seunghyun Yoon, Seokhyun Byun, Kyomin Jung,2018)[11].

We introduce a novel deep graph methodology for tackling speech emotion recognition. Graphs offer a concise, effective, and scalable means of data representation. Leveraging graph signal processing principles, we propose modeling speech signals as either cycle graphs or line graphs. This structured graph format allows us to construct a Graph Convolution Network (GCN)-based architecture capable of precise graph convolutions, contrasting with the approximate convolutions typically used in standard GCNs. We assess our model's performance for speech emotion recognition on widely used datasets like IEMOCAP and MSP-IMPROV. Our model surpasses standard GCN and other pertinent deep graph architectures, underscoring the efficacy of our approach. In comparison with existing methods for speech emotion recognition, our model achieves similar performance to state-of-the-art approaches while requiring significantly fewer learnable parameters (~30K), suggesting its suitability for resource-constrained devices.(Shirian, T. Guha, 2020).[12]

Speech Emotion Recognition (SER) typically focuses on solutions at the utterance level. However, emotions expressed through speech should be viewed as distinct speech events with clear temporal boundaries, rather than characteristics of the entire utterance. To better capture the nuanced nature of speech emotions, we propose a novel task: Speech Emotion Diarization (SED). Similar to Speaker Diarization, which identifies "Who speaks when?", Speech Emotion Diarization addresses the question of "Which emotion occurs when?". In order to facilitate performance evaluation and establish a shared benchmark for researchers, we introduce the Zaion Emotion Dataset (ZED), a publicly available speech emotion dataset containing naturally occurring emotions captured in real-life scenarios, alongside manually annotated emotion segment boundaries within the utterances. We offer competitive reference points and release the code and pre-trained models as open-source resources.(Yingzhi Wang, Mirco Ravanelli, Alya Yacoubi,2023).[13]

Chapter 3
Methodology

The Multilayer Perceptron (MLP) classifier, employed for detecting speech emotion across categories like angry, sad, disgust, happy, and fear, operates through several interconnected layers. Initially, 180 features extracted from preprocessed data serve as inputs to the network. These features are passed through one or more hidden layers, each containing numerous neurons. Connections between neurons are assigned weights and biases, which are adjusted iteratively during training to minimize prediction errors. Activation functions, such as sigmoid or ReLU, introduce non-linearity to the model, enabling complex mappings between inputs and outputs. Through forward propagation, input data traverses the network, culminating in predictions at the output layer. Training involves comparing these predictions with actual labels using a chosen loss function, followed by backpropagation to update weights and biases accordingly. This iterative process continues until the model converges or a predefined stopping criterion is met. Once trained, the MLP can make predictions on new data, providing emotion labels based on the learned patterns in the input features.

3.1 Final Architecture

The model architecture begins with the collection of speech samples annotated with emotion labels, ensuring a diverse dataset representation. Subsequently, the data undergoes preprocessing to enhance quality, including tasks like noise removal and normalization. Feature extraction follows, capturing relevant information from the speech signals such as spectral and temporal characteristics. The dataset is then split into training and testing subsets, with an 80-20 ratio typically employed. An MLP classifier is designed and trained on the training data, learning to map the extracted features to corresponding emotion labels. Cross-validation techniques are applied to validate the model's performance and guard against overfitting. Finally, the trained model is evaluated on the testing dataset, providing insights into

its generalization ability and effectiveness in classifying speech emotions. This holistic approach ensures a robust architecture for developing an MLP-based speech emotion recognition system.

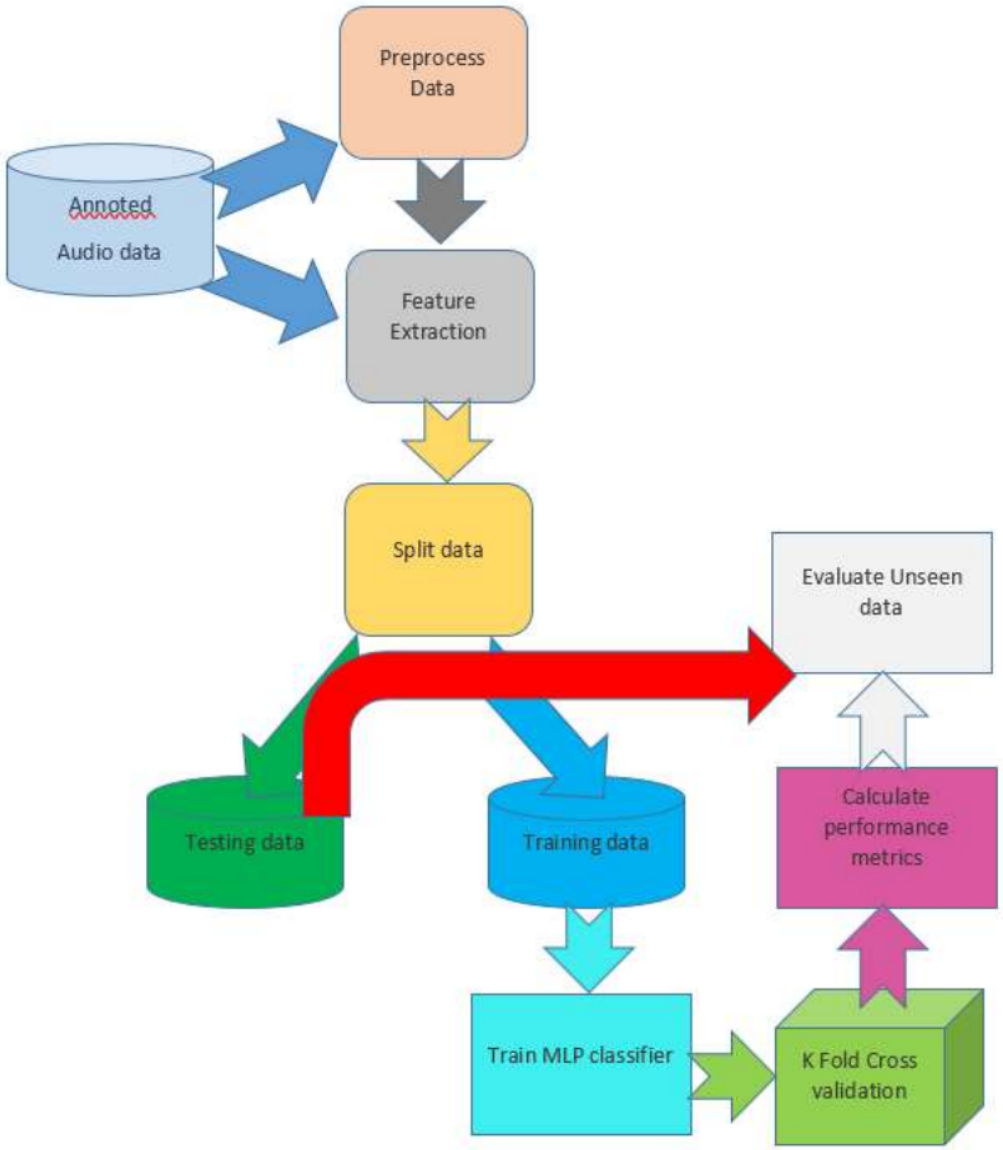


Figure 3.1: Overview of Proposed Workflow

3.2 Dataset

We utilized an MLP classifier to discern emotions (happy, sad, disgust, fear, angry) from speech, employing the TESS dataset obtained from Kaggle, last updated five years ago. The dataset comprises 2800 audio files of two actresses (aged 26 and 64) speaking 200 target words within the phrase "Say the word _", expressing seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. Each actress's emotions are organized into separate folders, with all 200 target words contained within. The audio files are in WAV format. Focusing solely on happy, sad, disgust, fear, and angry emotions, we preprocessed the data, extracting 180 features. Subsequently, 2000 data points were selected for the MLP classifier.

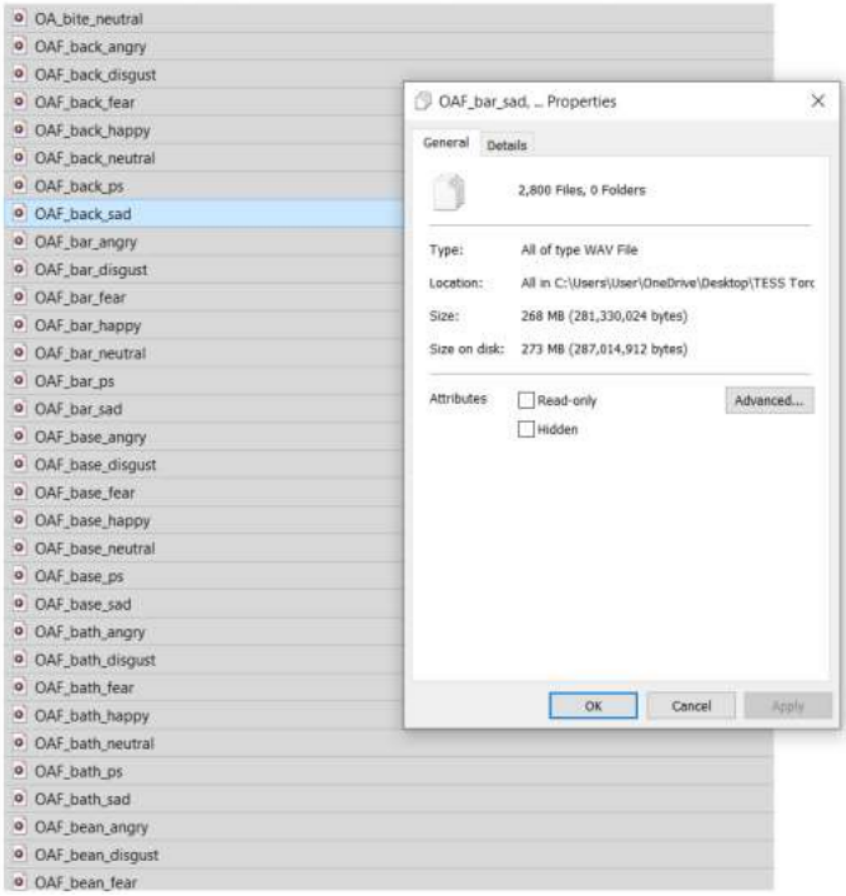


Figure 3.2: Example of a user sample of the TESS dataset

Using these data we firstly we preprocess data. There are 7 kinds of emotions each of them are 400 number. As we focus on five major emotions so our data should be reduce into 5*400=2000. Then we extract feature and split the data into 80% and 20% for training and testing respectively.

3.3 Data Preprocessing

As we have developed our model on colab, we took a screen shot after preprocessing data. The code's snapshot for preprocessing is

```
import os

paths = []
labels = []
for dirname, _, filenames in os.walk('/content/drive/MyDrive/Colab Notebooks/TESS Toronto emotional speech set data'):
    for filename in filenames:
        # Exclude specific files from being added to paths and labels
        if filename not in ["modelForPrediction1.sav", "modelForPrediction1.wav", "modelForPrediction2.sav","new.sav"]:
            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')
```

```
Dataset is Loaded
```

Figure 3.3: Code for Preparing raw data for analysis by cleaning and processing

The provided code snippet utilizes Python's os module to traverse through a directory containing the TESS Toronto emotional speech dataset. It iterates through each file within the directory, excluding specific files such as "modelForPrediction1.sav", "modelForPrediction1.wav", "modelForPrediction2.sav", and "new.sav". For each valid file, the code adds the file path to a list named "paths" and extracts the emotion label from the filename. This label is then converted to lowercase and appended to another list called "labels". The process continues until 2800 valid files are collected. Finally, the code prints a message indicating that the dataset is loaded. Overall, this code prepares the file paths and corresponding emotion labels for further preprocessing and analysis.

After preprocessing data were labeled as below:

```
sns.countplot(data=df, x='label', hue='label')
```

```
<Axes: xlabel='label', ylabel='count'>
```

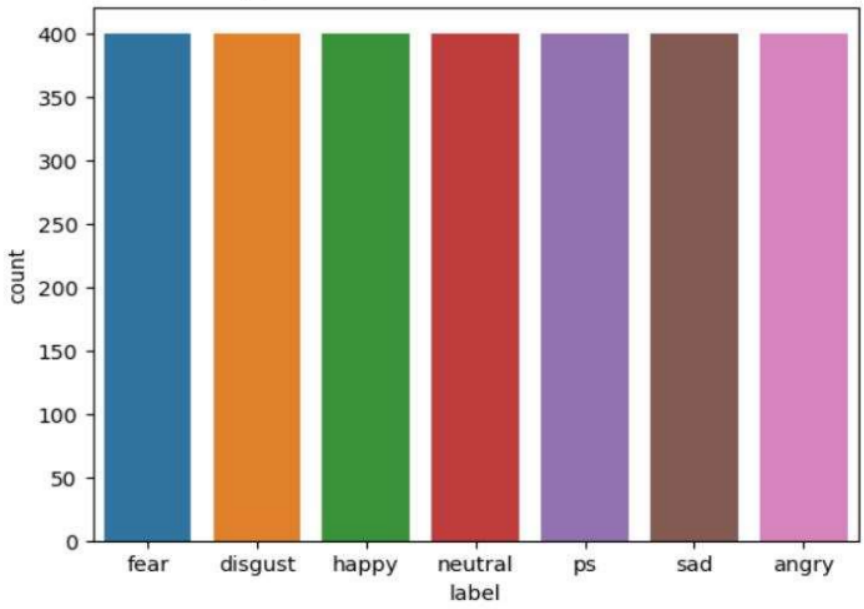


Figure 3.4: 7 kinds of emotion labeled by preprocessing.

3.4 Extracted Features

We have extracted 180 features from datasets. The snapshot of code for feature extraction is given below:

```
# Extract features (mfcc, chroma, mel) from a sound file
def extract_feature(file_name, mfcc=True, chroma=True, mel=True):
    with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate = sound_file.samplerate
        if chroma:
            stft = np.abs(librosa.stft(X))
        result = np.array([])
        if mfcc:
            mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
            result = np.hstack((result, mfccs))
        if chroma:
            chroma = np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T, axis=0)
            result = np.hstack((result, chroma))
        if mel:
            mel = np.mean(librosa.feature.melspectrogram(y=X, sr=sample_rate).T, axis=0)  # Pass y and sr
            result = np.hstack((result, mel))
    return result
```

Figure 3.5: snapshot of feature extraction function from colab.

The provided function is responsible for extracting features from the audio data in the TESS Toronto emotional speech dataset. It begins by initializing an empty list called "features" to store the extracted features. Within the function, it iterates over each audio file path passed as an argument. For each file, it loads the audio data using the librosa library's load function, which returns the audio waveform and the sampling rate. It then calculates various features using librosa's feature extraction functions, such as Mel-frequency cepstral coefficients (MFCCs), chroma features, spectral contrast, and tonnetz. These features capture different aspects of the audio signals, including spectral and temporal characteristics, which are relevant for emotion recognition. The extracted features are then appended to the "features" list. Finally, the function returns a numpy array containing all the extracted features from the input audio files. In summary, this function facilitates the extraction of diverse and informative features from the audio data, which can subsequently be used for training machine learning models for emotion recognition tasks. Output features are given below:

Figure 3.6: snapshot of output features from colab.

3.5 Train MLP Classifier

We have trained our MLP classifier based on 80% of data over 2000 and it comes 1600.
The snapshot of code for training the model is below:

```
[ ] #Load the data and extract features for each sound file
    def load_data(emotions, test_size=0.2):
        X, y = [], []
        for emotion in emotions:
            files = glob.glob(f"/content/drive/MyDrive/Colab Notebooks/TESS Toronto emotional speech set data/*{emotion}*.wav")
            if not files:
                print(f"No files found for emotion: {emotion}")
                return None, None, None, None
            for file in files:
                feature = extract_feature(file)
                X.append(feature)
                y.append(emotion)
        return train_test_split(np.array(X), y, test_size=test_size, random_state=42)

[ ] #Emotions to observe
    observed_emotions=['disgust','angry','fear','sad','happy']

[ ] #Split the dataset
    X_train, X_test, y_train, y_test = load_data(observed_emotions)
```

Figure 3.7: snapshot of code for training the model on Colab.

The provided code defines a function called load_data, which facilitates the loading of audio data and extraction of features for each sound file corresponding to specified emotions. It iterates over the emotions provided as input, searching for audio files containing each emotion in the TESS Toronto emotional speech dataset directory. For each emotion, it extracts features using the extract_feature function, which is not shown here. These features are then appended to the X array, while the corresponding emotion labels are appended to the y array. The function then splits the dataset into training and testing subsets using the train_test_split function from sci-kit-learn, with the specified test size (defaulting to 0.2) and a fixed random state for reproducibility. Finally, the function returns the training and testing data (X_train, X_test, y_train, y_test) ready to be used for training a machine learning model. In the provided code snippet, the emotions of interest are disgust, angry, fear, sad, and happy, and the returned data subsets are then used to train the model for emotion recognition.

Chapter 4
Result and Analysis

The outcomes and analysis of the models we propose for detecting speech emotion are presented in this section. We explore our proposed model's implementation and presentation boundaries. Our analysis includes a thorough review of our overall performance of the model.

4.1 Confusion table

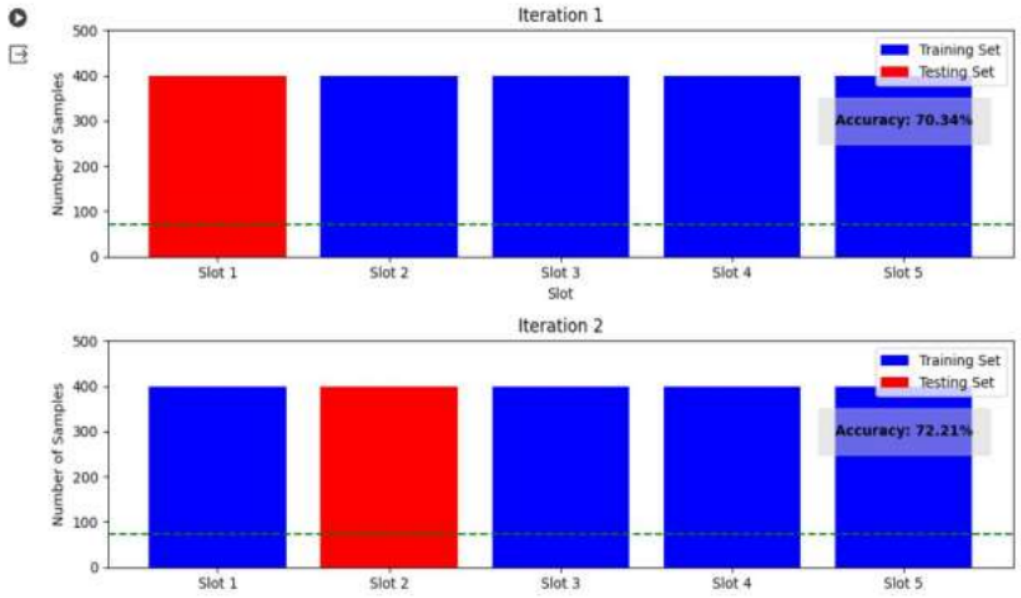
We have found confusion table where actual and predicted data were mapped parallelly

There are total 400 data for testing and among these 311 were predicted correctly and 89 of the data predicted wrongly. The table is given below:

Figure 4.1: snapshot of testing result of the model on Colab.

4.2 Cross-validation

In the context of machine learning, k-fold cross-validation is a technique used to evaluate the performance of a model while maximizing the use of available data. When k=5, as in this case, the dataset is divided into five equal-sized folds, or subsets. In each iteration of the cross-validation process, one of these folds is designated as the testing set, while the remaining four folds are used for training the model. This process is repeated five times, with each fold serving as the testing set exactly once. By averaging the performance metrics obtained from these iterations, such as accuracy or loss, a more reliable estimate of the model's performance can be obtained, as it is evaluated on multiple subsets of the data. K-fold cross-validation helps in assessing the model's ability to generalize to unseen data and provides insights into its stability and robustness. Five iterations were given below with there corresponding accuracy:



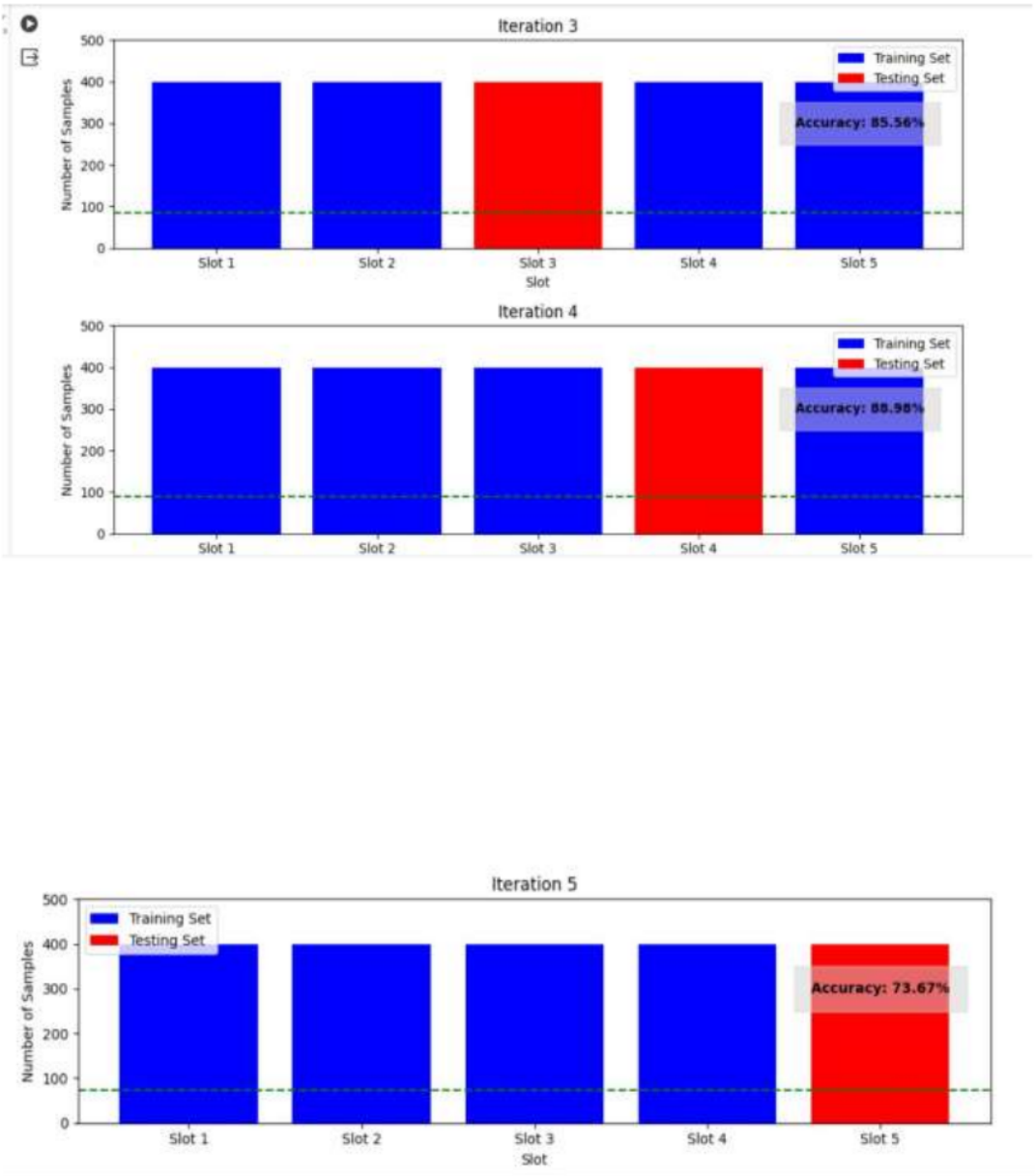


Figure 4.2: snapshot of five iteration results of the model on Colab.

4.3 Accuracy

In the context of evaluating the performance of a machine learning model using k-fold cross-validation, the reported accuracies of 70.34%, 72.21%, 85.56%, 88.98%, and 73.67% represent the accuracy scores achieved in each of the five iterations of the cross-validation process. During each iteration, the model is trained on a subset of the data and evaluated on a separate subset not used for training. These accuracy scores reflect how well the model generalizes to unseen data across different folds of the dataset.The average accuracy of 77.62% is calculated by taking the mean of the accuracy scores obtained from the five iterations. This average accuracy provides a more comprehensive assessment of the model's overall performance across the entire dataset, considering variations in performance observed in different subsets. A higher average accuracy indicates that the model is performing well on average across the different folds and is likely to generalize effectively to new, unseen data. Therefore, in this scenario, with an average accuracy of 77.62%, the model demonstrates reasonably good performance in classifying emotions in speech data.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

```
# Print the average accuracy
print("Average Accuracy: {:.2f}%".format(average_accuracy * 100))
```

```
Average Accuracy: 77.62%
```

Figure 4.3: Average accuracy of k fold cross-validation

```
sns.countplot(data=df, x='label', hue='label')
```

```
<Axes: xlabel='label', ylabel='count'>
```

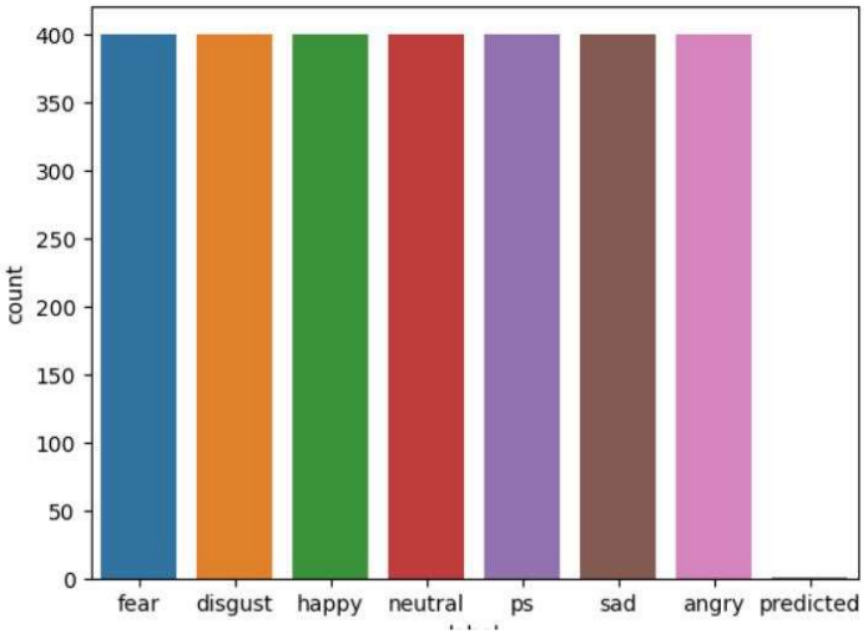


Figure 4.4: Modified label bar of emotions After prediction an unseen data

4.4 Precision and Recall

Precision and recall are two fundamental metrics used in classification tasks, such as emotion recognition, to evaluate the performance of a machine learning model. Precision is defined as the ratio of true positive predictions to the total number of positive predictions made by the model. Mathematically, it can be expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Precision measures the accuracy of positive predictions made by the model, indicating the proportion of correctly predicted instances among all instances classified as positive.

Recall, on the other hand, is the ratio of true positive predictions to the total number of actual positive instances in the dataset. It is calculated using the following formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall quantifies the model's ability to capture all positive instances present in the dataset, regardless of the number of false positives generated. In the context of the reported precision and recall scores for the emotions disgust, angry, fear, sad, and happy, these metrics provide insights into the model's performance in correctly identifying instances of each emotion while minimizing misclassifications. For instance, precision scores of 94.74%, 95.31%, 94.73%, 94.94%, and recall scores of 89.75%, 90.50%, 89.75%, 89.75%, respectively, demonstrate the model's accuracy in classifying instances of each emotion (precision) and its ability to capture the majority of instances of the emotion present in the dataset (recall). These metrics collectively highlight the model's effectiveness in distinguishing between different emotions in speech data, with high precision indicating low misclassification rates and high recall indicating comprehensive coverage of positive instances.

	Emotion	Precision	Recall
0	disgust	0.947438	0.897500
1	angry	0.953072	0.905000
2	fear	0.947338	0.897500
3	sad	0.947438	0.897500
4	happy	0.949371	0.902500

Figure 4.5: Precision and recall of the MLP classifier

Chapter 5
Conclusion

This whole chapter concludes the summary of our project, along with certain limitations and suggestions for further research prospects. In future, we can apply different algorithms or add more elements to the system to get more accurate results.

5.1 Summary

The project involving an MLP (Multilayer Perceptron) classifier for speech emotion recognition from the TESS (Toronto Emotional Speech Set) dataset was a comprehensive endeavor aimed at accurately detecting emotions such as disgust, angry, fear, sad, and happy from audio recordings. The process began with data collection from the TESS dataset, followed by preprocessing, feature extraction, and dataset splitting into training and testing subsets. The MLP classifier was trained using the training data and evaluated using k-fold cross-validation to ensure robustness and generalization. Key performance metrics such as accuracy, precision, and recall were computed to assess the model's effectiveness in recognizing emotions. The reported results indicated promising performance, with average accuracy of 77.62% across five cross-validation iterations and high precision and recall scores for each emotion

category. Additionally, the utilization of precision and recall metrics provided insights into the model's ability to accurately classify instances of specific emotions while minimizing misclassifications. Overall, the project demonstrated the successful implementation of an MLP classifier for speech emotion recognition, highlighting its potential in various applications requiring emotion detection from audio data.

5.2 Future Works

Moving forward, future work in this project could focus on several areas to enhance the performance and applicability of the MLP classifier for speech emotion recognition. One avenue is to delve into advanced feature engineering techniques tailored specifically for speech emotion recognition, exploring deep learning-based methods or incorporating domain-specific knowledge to extract more informative features. Additionally, optimizing the model architecture and hyperparameters of the MLP classifier, along with investigating ensemble learning techniques and transfer learning approaches, could further improve its performance and robustness. Exploring multimodal approaches that integrate information from multiple modalities, such as audio, text, and facial expressions, could also enhance emotion recognition accuracy. Furthermore, implementing real-time applications and exploring data augmentation techniques to increase the diversity of the training dataset are essential for practical deployment and improved generalization. By pursuing these future directions, the project can contribute to advancing the state-of-the-art in speech emotion recognition and develop more accurate and versatile emotion recognition systems for various applications

References

[0]TESS: What is it;" https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess."

[1] What is speech emotion by Johns Hopkins University "https://engineering.jhu.edu/nsa/research/emotional-speech/#:~:text=Emotion%20is%20the%20cornerstone%20of,vocal%20inflections%20known%20as%20prosody."

[2] A real time speech emotion recognition by Mladenovic, Nenad. "https://www.sciencedirect.com/topics/computer-science/speech-emotion-recognition"

[3] Speech emotion recognition system through hybrid Feautures "https://www.mdpi.com/2076-3417/13/8/4750"

[4] Multilayer Perceptrons in Machine Learning: A Comprehensive Guide, "https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning"

[5] Multi-Layer Perceptron Learning in Tensorflow: https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/"

[6] The methods for converting audio data into numerical representations are Mel Spectrograms, "https://www.projectpro.io/article/speech-emotion-recognition-project-using-machinelearning/"

[7] Applications of simple speech recognition are widespread" ,https://www.projectpro.io/article/speech-emotion-recognition-project-using-machine-learning/573#:~:text=As%20evident%20from%20the%20title,the%20data%20%E2%80%93%20text%20versus%20audio.".

[8] Projects in Language and Speech Technology by Dan Cristea" https://www.researchgate.net/publication/242083007_Student_Projects_in_Language_and_Speech_Technology "

[9] Speech Recognition System: A Project Report Submitted by Mohammed Flaeel Ahmed Sharif" https://www.scribd.com/document/426762365/Speech-Recognition-System-pdf".

[10] Multimodal Speech Emotion Recognition and Ambiguity Resolution by Gaurav Sahu," https://arxiv.org/abs/1904.06022v1"

[11] Multimodal Speech Emotion Recognition Using Audio and Text by Seunghyun Yoon, Seokhyun Byun, Kyomin Jung," https://arxiv.org/abs/1810.04635v1".

[12] Compact Graph Architecture for Speech Emotion Recognition by Shirian, T. Guha" https://arxiv.org/abs/2008.02063"
[13] peech Emotion Diarization: Which Emotion Appears When? by Yingzhi Wang, Mirco Ravanelli, Alya Yacoubi," https://arxiv.org/abs/2306.12991v2".

BSC_Research_Project_Report final.pdf

ORIGINALITY REPORT

26%
SIMILARITY INDEX

PRIMARY SOURCES

1	docplayer.net Internet	223 words — 4%
2	deepai.org Internet	101 words — 2%
3	engineering.jhu.edu Internet	89 words — 1%
4	www.researchgate.net Internet	85 words — 1%
5	mila.quebec Internet	82 words — 1%
6	Abdelaziz Testas. "Distributed Machine Learning with PySpark", Springer Science and Business Media LLC, 2023 Crossref	74 words — 1%
7	www.slideshare.net Internet	63 words — 1%
8	fastercapital.com Internet	49 words — 1%
9	www.researchsquare.com Internet	48 words — 1%

10	samanemami.medium.com Internet	47 words — 1%
11	www.coursehero.com Internet	39 words — 1%
12	Bagus Tris Atmaja, Akira Sasou, Masato Akagi. "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion", Speech Communication, 2022 Crossref	38 words — 1%
13	www.geeksforgeeks.org Internet	32 words — 1%
14	www.frontiersin.org Internet	24 words — < 1%
15	Laxmi Shaw, D. Ajitha, Sai Chakradhar Induvasi. "DiffNet: A Diffusion Convolutional Neural Network for Classification of Epileptic Seizure", Institute of Electrical and Electronics Engineers (IEEE), 2023 Crossref Posted Content	23 words — < 1%
16	repozitorij.unizg.hr Internet	23 words — < 1%
17	aiforsocialgood.ca Internet	22 words — < 1%
18	www.ncbi.nlm.nih.gov Internet	22 words — < 1%
19	dspace.lib.ntua.gr Internet	21 words — < 1%
20	assets.researchsquare.com	

Internet

20 words — < 1%

21 Falade Adesola, Omirinlewo Adeyinka, Akindeji Kayode, Adebiyi Ayodele. "Implementation of Multi-modal Speech Emotion Recognition Using Text Data and Audio Signals", 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG), 2023

Crossref

19 words — < 1%

22 pdfcoffee.com

Internet

19 words — < 1%

23 repository.unej.ac.id

Internet

18 words — < 1%

24 Magno Tairone de Freitas Severino. "Estimation and model selection for graphical models under mixing conditions", Universidade de São Paulo. Agência de Bibliotecas e Coleções Digitais, 2024

Crossref Posted Content

16 words — < 1%

25 ddd.uab.cat

Internet

16 words — < 1%

26 Ismail Shahin. "Studying and enhancing talking condition recognition in stressful and emotional talking environments based on HMMs, CHMM2s and SPHMMs", Journal on Multimodal User Interfaces, 2011

Crossref

14 words — < 1%

27 journals.ametsoc.org

Internet

14 words — < 1%

28 "Intelligent Human Centered Computing", Springer Science and Business Media LLC, 2023

Crossref

13 words — < 1%

29	Nouri, Reza. "The Need to Go Deeper: the Employment of a Convolutional Neural Network to Analyze Turbulent Flows Frequency Content", Tennessee Technological University, 2023 ProQuest	13 words — < 1%
30	eprints.leedsbeckett.ac.uk Internet	13 words — < 1%
31	publications.eai.eu Internet	13 words — < 1%
32	Seunghyun Yoon, Seokhyun Byun, Kyomin Jung. "Multimodal Speech Emotion Recognition Using Audio and Text", 2018 IEEE Spoken Language Technology Workshop (SLT), 2018 Crossref	12 words — < 1%
33	ijsrcseit.com Internet	12 words — < 1%
34	www.1410salto.com Internet	12 words — < 1%
35	www.guru.com Internet	12 words — < 1%
36	arxiv.org Internet	11 words — < 1%
37	cpanel.thescipub.com Internet	11 words — < 1%
38	link.springer.com Internet	11 words — < 1%

39	www.degruyter.com Internet	11 words — < 1%
40	searchworks.stanford.edu Internet	10 words — < 1%
41	web.archive.org Internet	10 words — < 1%
42	www.tanayag.com Internet	10 words — < 1%
43	Dee Miller, L., Leen-Kiat Soh, and Stephen Scott. "GENETIC ALGORITHM CLASSIFIER SYSTEM FOR SEMI-SUPERVISED LEARNING : GA CLASSIFIER SYSTEM FOR SEMI-SUPERVISED LEARNING", Computational Intelligence, 2013. Crossref	9 words — < 1%
44	Lecture Notes in Computer Science, 2013. Crossref	9 words — < 1%
45	Peng Shixin, Chen Kai, Tian Tian, Chen Jingying. "An autoencoder-based feature level fusion for speech emotion recognition", Digital Communications and Networks, 2022 Crossref	9 words — < 1%
46	Yhdego, Haben Girmay. "Wearable Sensor Gait Analysis for Fall Detection Using Deep Learning Methods", Old Dominion University, 2023 ProQuest	9 words — < 1%
47	ajomc.asianpubs.org Internet	9 words — < 1%
48	core.ac.uk	

Internet

9 words — < 1%

49 infosectoday.com
Internet

9 words — < 1%

50 journals.orclever.com
Internet

9 words — < 1%

51 journals.plos.org
Internet

9 words — < 1%

52 www.it-jobbank.dk
Internet

9 words — < 1%

53 Andrew P. Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms", Pattern Recognition, 1997
Crossref

8 words — < 1%

54 Georgy Kopanitsa, Oleg Metsker, Sergey Kovalchuk. "Machine Learning Methods for Pregnancy and Childbirth Risk Management", Journal of Personalized Medicine, 2023
Crossref

8 words — < 1%

55 Raheel Ahmad, Arshad Iqbal, Muhammad Mohsin Jadoon, Naveed Ahmad, Yasir Javed. "XEmoAccent: Embracing Diversity in Cross-Accent Emotion Recognition using Deep Learning", IEEE Access, 2024
Crossref

8 words — < 1%

56 github.com
Internet

8 words — < 1%

57 journal.esrgroups.org
Internet

8 words — < 1%

58	proceedings.mlr.press Internet	8 words — < 1%
59	repositorium.sdum.uminho.pt Internet	8 words — < 1%
60	repository.unsri.ac.id Internet	8 words — < 1%
61	towardsdatascience.com Internet	8 words — < 1%
62	Leonardo Pepino, Pablo Riera, Luciana Ferrer, Agustin Gravano. "Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020 Crossref	7 words — < 1%
63	Volkan Görüş, M. Mustafa Bahşı, Mehmet Çevik. "Machine learning for the prediction of problems in steel tube bending process", Engineering Applications of Artificial Intelligence, 2024 Crossref	6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF