St. Thomas' College of Engineering and Technology

Synopsis on

Emo-Track

Department of Information Technology

By

Name of the Students	University Roll No.	University Registration No.
Pradip Ghosh	12200220013	201220100210050
Deep Kumar Goenka	12200220038	201220100210025
Aishi Paul	12200220039	201220100210024
Abhishek Chaudhury	12200220056	201220100210007

Under the guidance of

Dr. Arijit Ghosal

St. Thomas' College of Engineering and Technology

Affiliated to

Maulana Abul Kalam Azad University of Technology, West Bengal

November, 2023

Declaration Page

We are submitting the synopsis on Emo-Track as a part of our final year seventh semester project under the guidance of Dr. Arijit Ghosal

Pradip Ghosh

Deep Kumar Goenka

Aishi Paul

Abhishek Chaudhary

Guide's Signature (for approval)

Dr. Arijit Ghosal

Departmental Stamp

Vision: To promote the advancement of learning in Information Technology through research oriented dissemination of knowledge which will lead to innovative applications of information in industry and society.

Mission: To incubate students grow into industry ready professionals, proficient research scholars and enterprising entrepreneurs.

To create a learner- centric environment that motivates the students in adopting emerging technologies of the rapidly changing information society.

To promote social, environmental and technological responsiveness among the members of the faculty and student.

PEO: PEO1: Excel in professional career, higher education and research.

PEO2: Demonstrate professionalism, entrepreneurship, ethical behavior, communication skills and collaborative team work to adapt the emerging trends by engaging in lifelong learning.

PEO3: Exhibit the skills and knowledge required to design, develop and implement IT solutions for real life problems.

Project Mapping with Program Outcomes

PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
2	2	3	2	2	-	-	2	2	2	2	3

Enter correlation levels 1, 2 or 3 as defined below:

1: Slight (Low) 2: Moderate (Medium) 3: Substantial (High)

Justification:

PO1 (Engineering Knowledge): The engineering knowledge has been applied here for the development of this project.

PO2 (Problem Analysis): A lot of identification and review upon similar projects have been taken under consideration to analyze the problem.

PO3 (Design/Development of solutions): Solution has been designed that meet the specified needs with appropriate consideration for the public health, safety, cultural, societal and environment.

PO4 (Conduct investigations of complex problems): Used research-based knowledge and research methods to provide valid conclusions.

PO5 (Modern Tool Usage): The whole project is based on modern tool usage and all the appropriate IT tools and techniques required for this project has been used so to achieve the most appropriate solution.

PO8 (Ethics): This project is completely ethical and do not contain any private content piracy.

PO9 (Individual and team work): This is a team based project and every individual has contributed to make the project a successful one.

PO10 (Communication): Communication with engineering community, writing effective reports and design documentation, making effective presentation has ben done.

PO11 (Project Management and Finance): Project management has been done quite well, maintaining the cost and efficiency to complete the project.

PO12 (Life-long learning): Recognized the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PSO: PSO1 (Professional Competency): Apply their knowledge in the field of information technology and contribute significantly to the corporate world by way of providing appropriate solutions to engineering problems and establish their skills in high performance computing, software engineering, programming and thrust areas like security and machine intelligence.

PSO2 (Academic Aptitude): Demonstrate their proficiency in analytical and critical thinking, methodologies of practical design, data analysis and interpretation through their technical expertise which will help them to excel in higher studies within the country and abroad.

Project Mapping with Program Specific Outcomes

PSO1	PSO2
3	2

Enter correlation levels 1, 2 or 3 as defined below:

1: Slight (Low) **2: Moderate (Medium)** **3: Substantial (High)**

Justification:

PSO1 (Professional Competency): Knowledge of Machine Learning has been applied to build an effective solution.

PSO2 (Academic Aptitude): Demonstrated the proficiency of critical thinking to solve the project which will help to excel in higher studies within the country and abroad.

Index Page

Table of Contents Page No.

1. Introduction

In the age of human-computer interaction, understanding and interpreting human emotions through speech is a pivotal aspect of creating empathetic and responsive technology. Emo-Track, a Speech Emotion Recognition project, endeavors to unravel the emotional undertones embedded in speech using advanced machine learning techniques. The project harnesses the power of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), a comprehensive dataset comprising 1440 audio files. These files encapsulate a spectrum of eight emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised. The dataset's richness is accentuated by variations in emotional intensity, with 192 audio samples for each emotion, except for Neutral, which has 96 samples.

Emo-Track's methodology involves feature extraction, focusing on the mean and standard deviation of Mel-Frequency Cepstral Coefficients (MFCC), Modulation Spectrogram features, and Zero Crossing Rate (ZCR). This results in a feature matrix of dimensions (1440, 592). Dimensionality reduction is then employed through Principal Component Analysis (PCA), reducing the feature space to (1440, 80). Standardization using StandardScaler ensures the compatibility of features for subsequent modeling.

The project adopts a robust approach to model training and evaluation. Support Vector Machines (SVM), a powerful classification algorithm, is employed with a radial basis function (RBF) kernel and tuned hyperparameters (C=3.0, gamma='auto'). To further enhance model reliability, K-Fold cross-validation with 20 splits is applied, providing a more nuanced understanding of the model's performance.

Emo-Track stands at the forefront of emotion-aware technology, paving the way for applications in human-computer interaction, sentiment analysis, and beyond. Through its innovative approach, Emo-Track exemplifies the potential of artificial intelligence to comprehend and respond to the intricacies of human emotion in the realm of speech.

1.1. Problem Statement

Analyzing human voice and recognizing eight types of emotions viz. Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised.

1.2. Problem Definition

The project aims to develop an Emotion Recognition system, named Emo-Track, focused on analyzing human voice to accurately identify and categorize eight distinct emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised. Leveraging the RAVDESS dataset, consisting of 1440 audio files with varying emotional expressions, the system employs advanced signal processing and machine learning techniques to extract meaningful features from the audio signals.

1.3. Objective

Emo-Track helps in recognizing emotions from human voice. Using this model, one will be able to:

- Understand the emotions of the speaker.
- Continue conversation with them according to their emotions.
- Increase businesses by analyzing their customers responses.
- Monitor mental health of human beings.
- Make the autistic children understand the emotions of the speaker.

1.4. Literature Survey / Background Study

For the project namely ‘Emo-Track’ which can recognize Emotions from Speech, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is used which is a benchmarked dataset made by Steven R. Livingstone (2018). Emotion Speech Audio Dataset is used which contains 1440 files. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression [1]. Beth Logan (2000) presented the article about Mel Frequency Cepstral Coefficients (MFCCs) for music modeling which discusses the use of MFCCs in speech recognition and examines their applicability to music and also investigates the two main assumptions of the MFCC process: the Mel frequency scale and the Discrete Cosine Transform (DCT) where the author find that the Mel scale is not harmful for music modeling and that the DCT is an appropriate transform for decorrelating music log spectra [2].

Brian E.D. Kingsbury et al. (1998) discussed about the robustness of speech recognition using modulation spectrograms which emphasizes the low-frequency modulations in speech and makes more resistant to noise and reverberation [3]. Fabien Gouyon et al. (2000) discussed an article on the Zero Crossing Rate (ZCR) for classification of percussive sounds which is computed over the decay regions of the sound [4]. The survey presented in El Ayadi et al. (2011) addressed speech emotion classification with a focus on three key components of SER including feature selection, speech classification schemes, and preparation of speech emotion databases. The authors discussed issues associated with the feature extraction step in SER, and presented different categories of features including global and local and continuous and voice quality features and explained the influence of each category on SER performance [5].

Koolagudi and Rao (2012) presented an SER survey that covers databases, features, and classifiers. They discussed the benefits of using different databases including elicited, acted, and natural databases. They also covered different classification techniques and categories of features including prosodic, excitation, and vocal-tract features. However, they do not capture the techniques to deal with the speaker dependency issues and do not discuss emerging ML techniques used in SER [6]. Anagnostopoulos et al. (2015)

presented a survey paper that investigate emotion recognition from audio channels and classify the papers based on the features extracted and selected for training the classifiers (linguistic or non-linguistic) and their classification methodology and provided some conclusions from the study, such as the need for further exploration of feature sets and classification methods, the need for a "golden set" of non-linguistic features, and the difficulty of developing a complete common database [7].

Schuller (2018) discussed the challenges involved in SER, such as the lack of data and the difficulty of labeling emotions despite which SER has made significant progress in recent years, and it is now being used in a variety of applications, such as customer service and healthcare [8]. Swain et al. (2018) presented an SER survey that captures components such as databases, features, and different classifiers for speech emotion recognition in several languages. Deep learning, hybrid, and fusion techniques for emotion classification were also discussed in their research. However, their survey did not discuss challenges and the current state-of-the-art approaches for achieving speaker independence, feature selection, evaluation, and several key techniques for speech data pre-processing [9].

Mustafa et al. (2018) discusses the article reviews existing research in SER and finds that it is an active field of research. The authors identify several key issues that are still not fully resolved, such as the use of more appropriate databases, research on under-resourced languages, real-time recognition, and the inclusion of new types of emotion. It provides a comprehensive overview of the current state of research and identifies areas in which further research is needed. The authors’ suggestions for future research are well-considered and could help to advance the field of SER in the years to come [10]. The survey conducted by Akçay and Oğuz (2020) reviews different methods for speech emotion recognition (SER) and discuss discrete and dimensional models of emotion and also discusses different techniques for speech data preprocessing and various ML-based and deep learning classifiers [11].

The article presented by Latif et al. (2021) discusses the progress of deep representation learning in the field of SER which emphasize that deep representation learning is very important in SER. The article also highlights some important points about deep representation learning in SER, such as the need for natural emotional corpora, the need for systems that use raw speech or input features, the popularity of LSTM/GRU-RNNs and CNNs, the potential of Transformers, the need for exploration of static deep representation learning methods, the need for privacy-preserving representation learning, and the potential of multimodal self-supervised domain adaptive representation learning models [12].

The survey paper discussed by Yadav et al. (2021) discusses the latest developments in machine learning architectures, algorithms, and applications for speech and vision systems and also discussed the challenges and successes of using machine learning on platforms with limited resources. They conclude by highlighting the promise of emerging speech and vision systems applications, such as efficient evaluation, and accurate medical prescriptions [13]. Fahad et al. (2021) discussed a survey paper on speech emotion recognition (SER) in natural environments. It discusses SER techniques for natural environments along with their advantages and disadvantages in terms of speaker, text, language, and recording environments. It also discusses deep learning techniques for SER, which have become popular in recent years due to their minimal speech processing and enhanced accuracy. Finally, the paper discusses recent databases, features, and feature selection algorithms for SER, which have not been discussed in previous surveys and can be promising for SER in natural environments [14].

1.5. Brief Discussion on Problem

The task of analyzing human voice and recognizing emotions falls under the domain of affective computing and speech emotion recognition. This challenging problem involves developing algorithms and models capable of extracting relevant features from audio signals to accurately classify and identify the underlying emotional states of individuals. Emotion recognition from speech has applications in various fields, including human-computer interaction, customer service, and mental health monitoring. Techniques often used for this task include signal processing, machine learning, and deep learning methods. Researchers typically rely on large datasets of labeled emotional speech to train and evaluate models, with the goal of achieving high accuracy and robustness across different speakers and linguistic variations. Success in this area could contribute to the development of more emotionally intelligent systems that can better understand and respond to human communication, enhancing the overall quality of human-machine interactions.

Emotions are fundamental components of human communication, expressed through various channels such as facial expressions, body movements, and verbal communication. Recognizing the significance of emotions in human interactions, researchers are increasingly adopting approaches that leverage audio signals for emotion acknowledgment. This paradigm shift has become particularly relevant in today's context, where the recognition of emotions plays a pivotal role in fostering effective and meaningful communication.

In the realm of emotion recognition, the focus has expanded to encompass diverse modalities such as body language, facial expressions, and voice recognition. While facial expressions and body language readily convey emotions in face-to-face conversations, the challenge arises when communication occurs through mediums that lack visual cues. In such scenarios, the medium itself becomes a channel for conveying emotions, making it challenging to predict the emotional state of the individual. Herein lies the importance of Speech Emotion Recognition (SER), a method that enables the expression of one's emotional state through spoken words.

Human vocal sounds, characterized by modulations in pitch, loudness, tone, and timbre, serve as distinct features that differentiate us from other living beings. Leveraging these vocal attributes, researchers can analyze and interpret human emotions effectively. The universality of certain emotions, such as anger, sadness, happiness, surprise, fear, and neutrality, makes them accessible for recognition by systems trained to identify these emotional states.

Feature extraction from human audio signals is a critical step in the process of emotion recognition. By extracting relevant information from the voice, such as pitch variations and tonal qualities, systems can accurately identify and categorize emotions. This not only facilitates more nuanced human-computer interactions but also holds significant implications for supporting individuals with physical disabilities who may face challenges in expressing their emotions through conventional means.

Beyond the realm of interpersonal communication, emotion recognition has broader implications, contributing to the well-being and accessibility of individuals who face difficulties in conventional communication channels. For instance, physically disabled individuals who may lack the means to convey emotions through traditional methods can benefit significantly from systems that can discern and interpret emotions through speech. As technology advances, the integration of emotion recognition not only enhances the efficiency of human-computer interactions but also promotes inclusivity and understanding in diverse communication contexts. Emotion recognition is thus a multifaceted tool with the potential to bridge gaps in communication and foster greater empathy and connection between individuals.

This system has the following unique features:

- Human-Computer Interaction (HCI): Enhances interactions with virtual assistants by adapting responses based on user emotions.

- Assistive Technology: Supports physically disabled individuals in expressing emotions through speech.

- Call Center Optimization: Analyzes customer sentiment in real-time for improved service and satisfaction.

- Emotion-aware Educational Tools: Integrates into language learning apps for feedback on pronunciation and emotional expression.

- Mental Health Monitoring: Aids in early detection of emotional distress or changes in mental health through speech analysis.

- Market Research: Provides sentiment analysis in market research for deeper consumer insights.

- Voice-based Emotion Analytics Platforms: Forms a core component in platforms for emotion analytics in marketing, entertainment, and healthcare.

1.5.1 Challenges

- Ambiguity in Expression: Emotions exhibit diverse expressions, posing challenges in accurate classification as individuals may convey the same emotion differently.

- Cultural and Individual Variability: Variations in emotional expression due to cultural and individual differences impact model generalization, hindering consistent performance across diverse speakers.

- Data Imbalance: Class imbalances in emotion datasets lead to biased models, excelling in overrepresented emotions but struggling with underrepresented ones.

- Real-time Processing Challenges: Implementing real-time systems demands low-latency processing, particularly challenging for applications like human-computer interaction or virtual assistants.

2. Concepts and Problem Analysis

2.1 Target User

This versatile model caters to researchers, AI practitioners, developers in human-computer interaction, customer service professionals, educators, physically disabled individuals, and industries in market research, entertainment, and gaming. It enhances user experiences, fosters emotionally-aware learning environments, aids sentiment analysis in real-time, and provides a unique avenue for expressing emotions through speech for those facing physical challenges.

2.2 Hardware and Software requirements:

2.2.1 Hardware Requirements:

- Operating System: Windows / Linux / macOS
- CPU: A multicore processor (AMD Ryzen or Intel i3 or above)
- RAM: 4 GB or more
- Storage: SSD with at least 4 GB free space

2.2.2 Software Requirements:

- Python: Version 3.x
- Libraries:
  - Librosa: Audio processing
  - NumPy, SciPy: Numerical operations
  - Scikit-learn: Machine learning tasks
  - Matplotlib, Seaborn: Data visualization
  - joblib: Model saving/loading
- Environment:
  - Google Colab or Jupyter Notebook
  - IDE: VSCode, PyCharm
  - Conda or Virtual Environment

2.3 Flow Chart

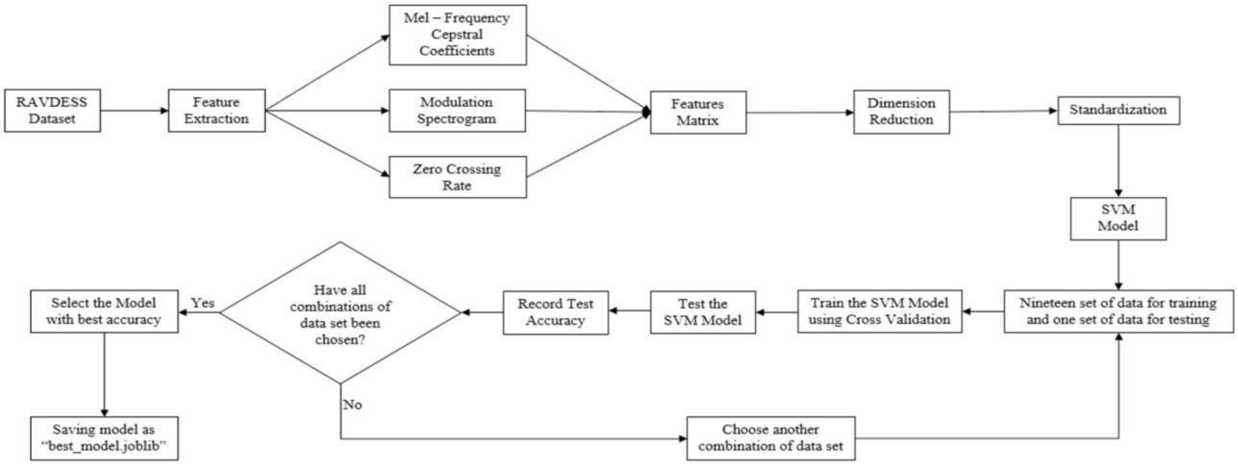

RAVDESS Dataset
Feature Extraction
Mel – Frequency Cepstral Coefficients
Modulation Spectrogram
Zero Crossing Rate
Features Matrix
Dimension Reduction
Standardization
SVM Model
Nineteen set of data for training and one set of data for testing
Train the SVM Model using Cross Validation
Test the SVM Model
Record Test Accuracy
Have all combinations of data set been chosen?
Yes
Select the Model with best accuracy
Saving model as “best_model.joblib”
No
Choose another combination of data set

Fig – 1: Flowchart of training and testing of Emo-Track model

This is the entire flowchart shown in Fig – 1 about the training process of the Emo-Track model which is trained on the RAVDESS Dataset containing 1440 audio files, extracting the acoustic features such as Mel – Frequency Cepstral Coefficients (MFCCs), Modulation Spectrogram and Zero Crossing Rate (ZCR) and generating a Feature matrix having dimension of (1440, 592). Now, PCA is applied for dimension reduction and reduced the dimension to (1440, 80). The feature matrix is standardized using StandardScalar. Support Vector Machines (SVM) Model is used to train the model and using KFold Cross Validation technique of 20 folds to get the model with the highest accuracy where nineteen folds represent the Training dataset and one set represents Testing dataset. After training, the model with the highest accuracy is getting saved into a joblib file named “best_model.joblib” and the highest accuracy is noted as 95.83%.

2.4 Gantt Chart

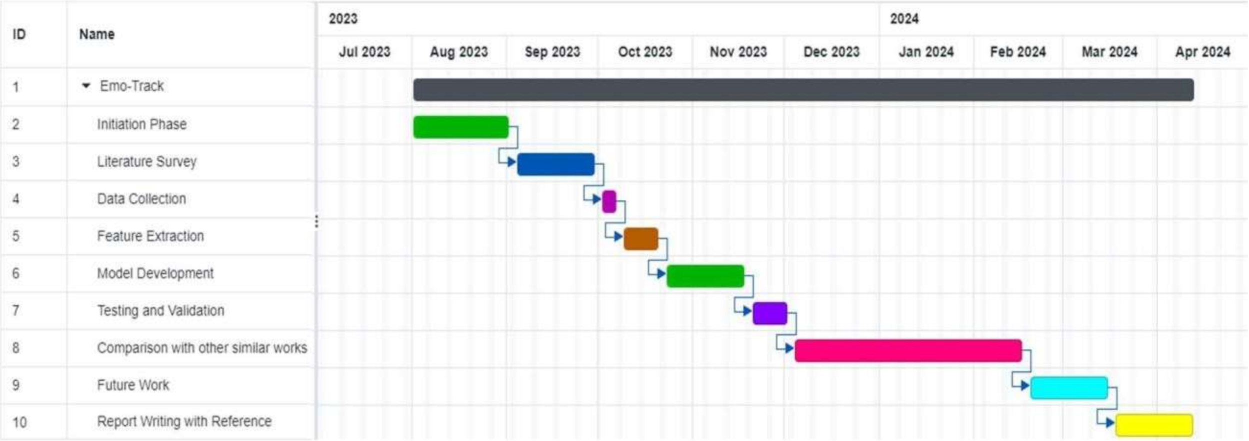


Fig – 2: Gantt Chart of Emo-Track model

Fig – 2 illustrates the Gantt Chart for the Emo-Track project, indicating the completion of the Testing and Validation stage. The remaining stages are scheduled to be finished by April 2024.

2.5 Dataset

The RAVDESS dataset, developed by Ryerson University, offers a robust foundation for Speech Emotion Recognition (SER). With 1440 audio files from 24 professional actors simulating eight emotions, including Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised, it features a unique blend of emotional speech. Notably, emotion samples vary, with 192 for most categories and 96 for Neutral where the count of emotions is shown as a bar chart in Fig – 3. The dataset's 16 kHz sampling rate ensures high-quality audio. Publicly accessible, RAVDESS supports collaborative research, yet ethical considerations regarding simulated emotional expressions are vital. While it provides a valuable resource for speech emotion recognition, researchers should be mindful of its controlled setting and potential limitations in capturing real-world contextual variations.

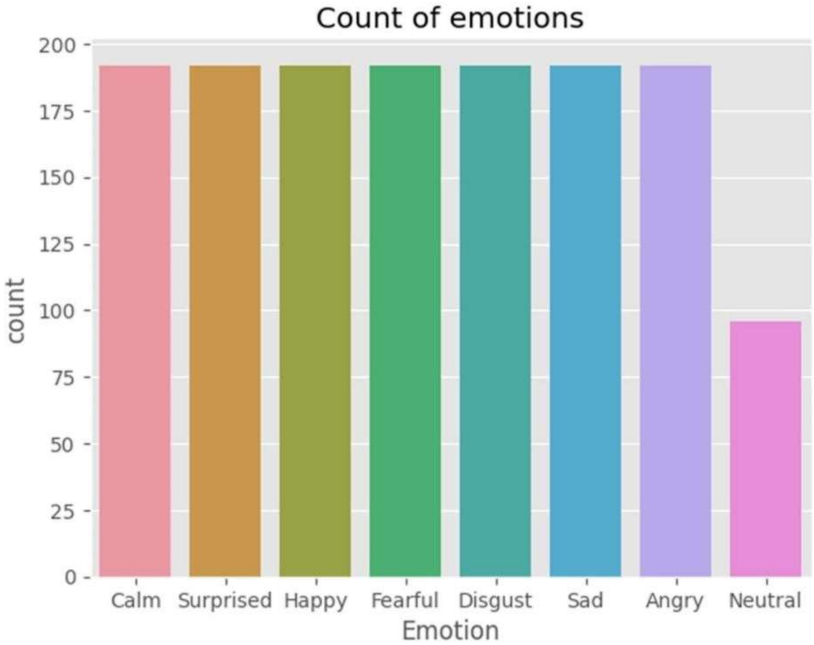


Fig – 3: Count of Emotion

2.6 Feature Extraction

In speech emotion recognition, three key feature extraction techniques are employed to capture diverse aspects of emotional content. Mel-frequency cepstral coefficients (MFCC) with n mfcc set at 39 provide insights into spectral characteristics, crucial for discerning emotional nuances. The modulation spectrogram, utilizing n_mels at 256, captures dynamic frequency changes over time. Extracting mean and standard deviation from both MFCC and the modulation spectrogram refines the representation, encompassing central tendencies and variability. Additionally, the zero-crossing rate (ZCR) serves to measure temporal characteristics, with mean and standard deviation contributing to a comprehensive understanding of speech signal transitions. This combination of features, each offering unique insights into the emotional expressiveness of speech, provides a nuanced and comprehensive foundation for subsequent stages in the speech emotion recognition pipeline.

2.7 Dimensionality Reduction

Principal Component Analysis (PCA), serves as a pivotal dimensionality reduction method in this scenario, transforming the original features into a set of 80 uncorrelated variables, known as principal components. These components are prioritized based on their ability to explain variance in the data, providing a condensed yet meaningful representation. The essence of this reduction lies in its practical applications: it compresses features, reduces noise by emphasizing crucial components, and accelerates machine learning efficiency through lower-dimensional computations. PCA optimizes data analysis and enhances machine learning tasks by preserving essential information in a more manageable format. Fig – 4 shows the Explained Variance Ratio Plot.

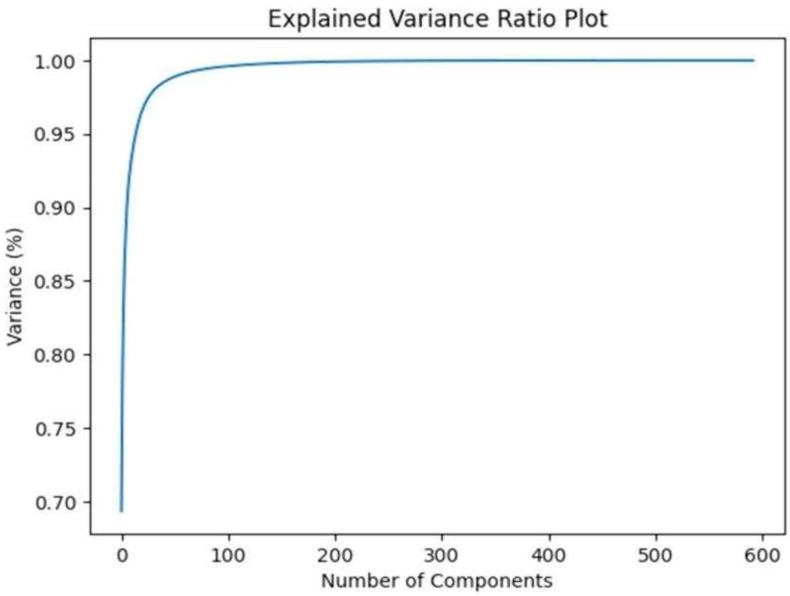
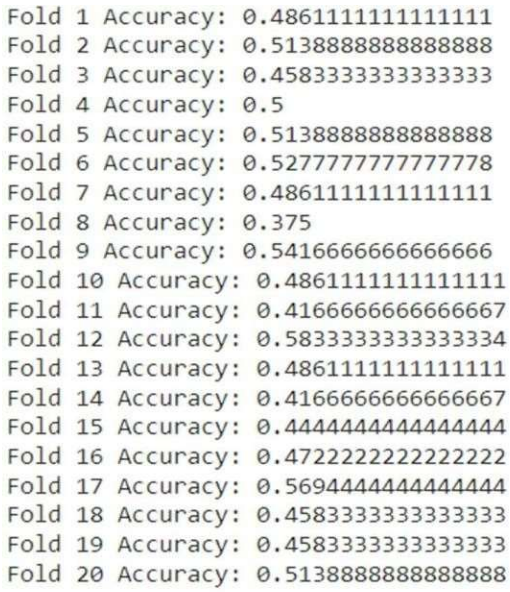


Fig – 4: Explained Variance Ratio Plot

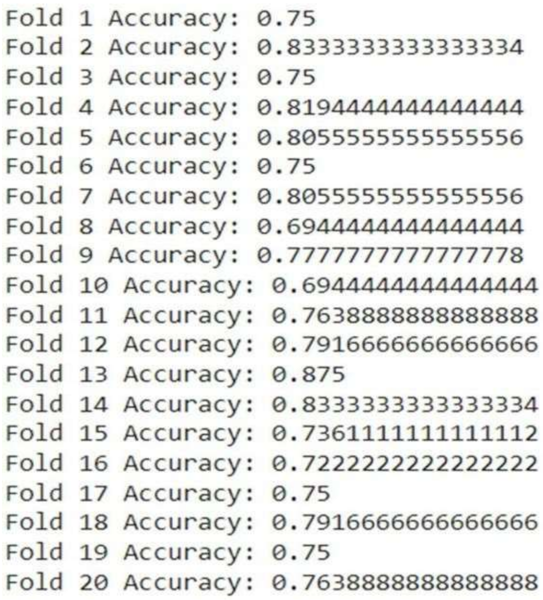
We conducted experiments using six different sets of principal components, specifically with dimensions 30, 40, 50, 60, 70, and 80. Among these, the set with 80 principal components yielded the highest accuracy of 95.83%. The detailed results of all the experiments are provided below:

Principal components = 30



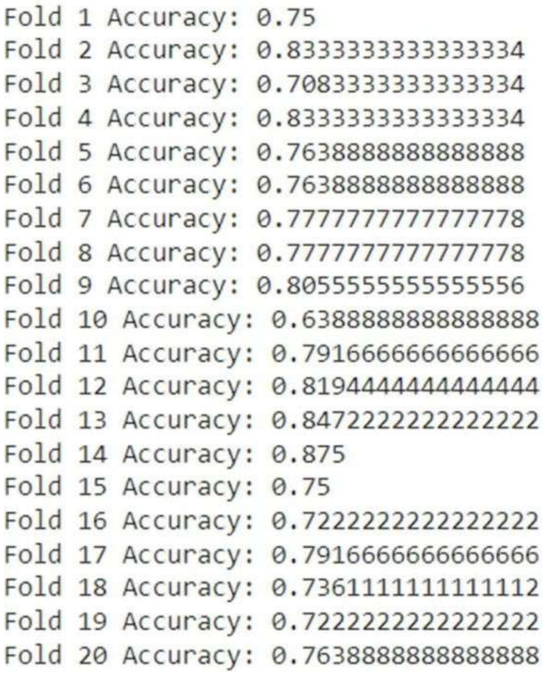
```
Fold 1 Accuracy: 0.4861111111111111
Fold 2 Accuracy: 0.5138888888888888
Fold 3 Accuracy: 0.4583333333333333
Fold 4 Accuracy: 0.5
Fold 5 Accuracy: 0.5138888888888888
Fold 6 Accuracy: 0.5277777777777778
Fold 7 Accuracy: 0.4861111111111111
Fold 8 Accuracy: 0.375
Fold 9 Accuracy: 0.5416666666666666
Fold 10 Accuracy: 0.4861111111111111
Fold 11 Accuracy: 0.4166666666666667
Fold 12 Accuracy: 0.5833333333333334
Fold 13 Accuracy: 0.4861111111111111
Fold 14 Accuracy: 0.4166666666666667
Fold 15 Accuracy: 0.4444444444444444
Fold 16 Accuracy: 0.4722222222222222
Fold 17 Accuracy: 0.5694444444444444
Fold 18 Accuracy: 0.4583333333333333
Fold 19 Accuracy: 0.4583333333333333
Fold 20 Accuracy: 0.5138888888888888
```

Principal components = 40



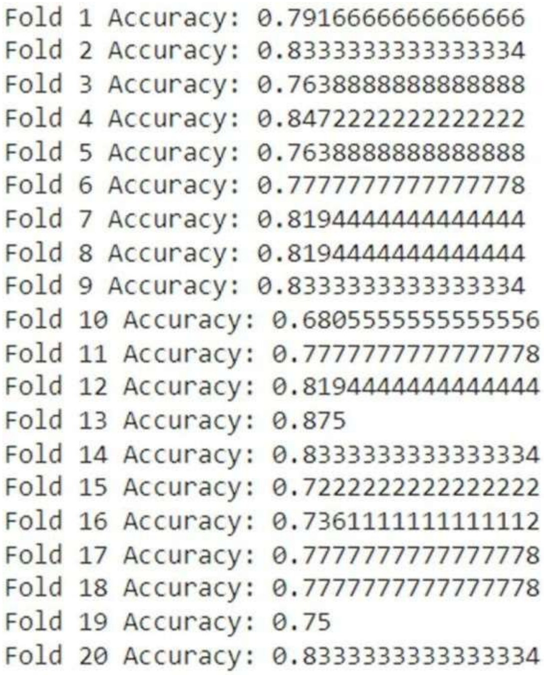
```
Fold 1 Accuracy: 0.75
Fold 2 Accuracy: 0.8333333333333334
Fold 3 Accuracy: 0.75
Fold 4 Accuracy: 0.8194444444444444
Fold 5 Accuracy: 0.8055555555555556
Fold 6 Accuracy: 0.75
Fold 7 Accuracy: 0.8055555555555556
Fold 8 Accuracy: 0.6944444444444444
Fold 9 Accuracy: 0.7777777777777778
Fold 10 Accuracy: 0.6944444444444444
Fold 11 Accuracy: 0.7638888888888888
Fold 12 Accuracy: 0.7916666666666666
Fold 13 Accuracy: 0.875
Fold 14 Accuracy: 0.8333333333333334
Fold 15 Accuracy: 0.7361111111111112
Fold 16 Accuracy: 0.7222222222222222
Fold 17 Accuracy: 0.75
Fold 18 Accuracy: 0.7916666666666666
Fold 19 Accuracy: 0.75
Fold 20 Accuracy: 0.7638888888888888
```

Principal components = 50



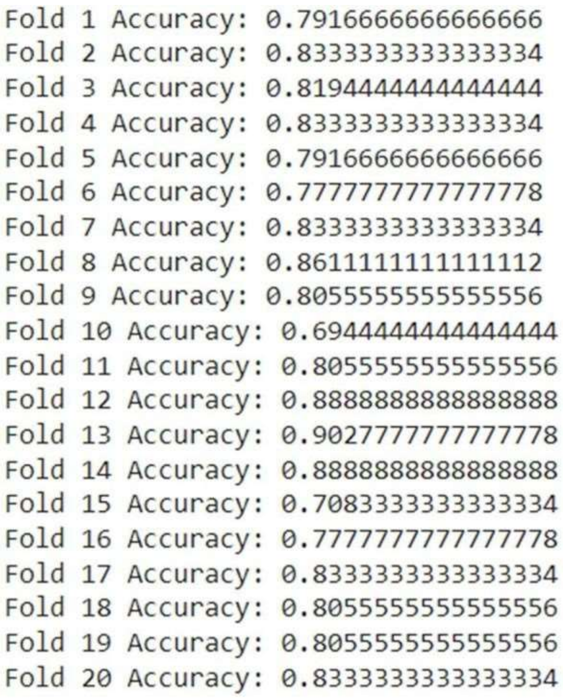
```
Fold 1 Accuracy: 0.75
Fold 2 Accuracy: 0.8333333333333334
Fold 3 Accuracy: 0.7083333333333334
Fold 4 Accuracy: 0.8333333333333334
Fold 5 Accuracy: 0.7638888888888888
Fold 6 Accuracy: 0.7638888888888888
Fold 7 Accuracy: 0.7777777777777778
Fold 8 Accuracy: 0.7777777777777778
Fold 9 Accuracy: 0.8055555555555556
Fold 10 Accuracy: 0.6388888888888888
Fold 11 Accuracy: 0.7916666666666666
Fold 12 Accuracy: 0.8194444444444444
Fold 13 Accuracy: 0.8472222222222222
Fold 14 Accuracy: 0.875
Fold 15 Accuracy: 0.75
Fold 16 Accuracy: 0.7222222222222222
Fold 17 Accuracy: 0.7916666666666666
Fold 18 Accuracy: 0.7361111111111112
Fold 19 Accuracy: 0.7222222222222222
Fold 20 Accuracy: 0.7638888888888888
```

Principal components = 60



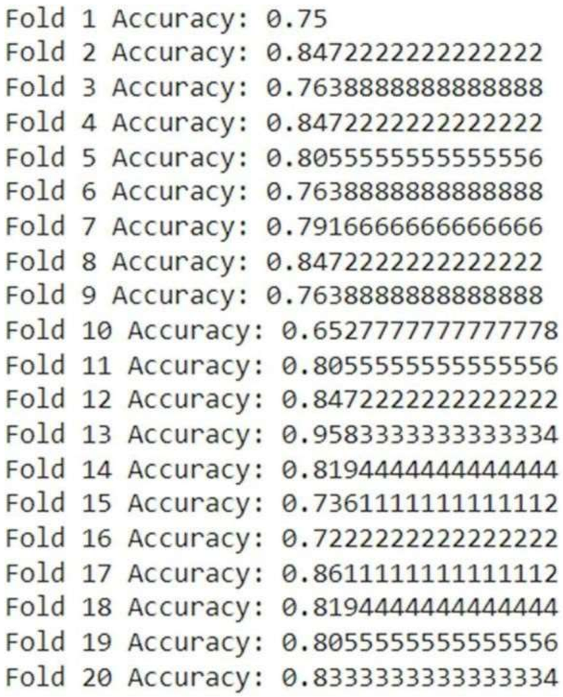
```
Fold 1 Accuracy: 0.7916666666666666
Fold 2 Accuracy: 0.8333333333333334
Fold 3 Accuracy: 0.7638888888888888
Fold 4 Accuracy: 0.8472222222222222
Fold 5 Accuracy: 0.7638888888888888
Fold 6 Accuracy: 0.7777777777777778
Fold 7 Accuracy: 0.8194444444444444
Fold 8 Accuracy: 0.8194444444444444
Fold 9 Accuracy: 0.8333333333333334
Fold 10 Accuracy: 0.6805555555555556
Fold 11 Accuracy: 0.7777777777777778
Fold 12 Accuracy: 0.8194444444444444
Fold 13 Accuracy: 0.875
Fold 14 Accuracy: 0.8333333333333334
Fold 15 Accuracy: 0.7222222222222222
Fold 16 Accuracy: 0.7361111111111112
Fold 17 Accuracy: 0.7777777777777778
Fold 18 Accuracy: 0.7777777777777778
Fold 19 Accuracy: 0.75
Fold 20 Accuracy: 0.8333333333333334
```

Principal components = 70



```
Fold 1 Accuracy: 0.7916666666666666
Fold 2 Accuracy: 0.8333333333333334
Fold 3 Accuracy: 0.8194444444444444
Fold 4 Accuracy: 0.8333333333333334
Fold 5 Accuracy: 0.7916666666666666
Fold 6 Accuracy: 0.7777777777777778
Fold 7 Accuracy: 0.8333333333333334
Fold 8 Accuracy: 0.8611111111111112
Fold 9 Accuracy: 0.8055555555555556
Fold 10 Accuracy: 0.6944444444444444
Fold 11 Accuracy: 0.8055555555555556
Fold 12 Accuracy: 0.8888888888888888
Fold 13 Accuracy: 0.9027777777777778
Fold 14 Accuracy: 0.8888888888888888
Fold 15 Accuracy: 0.7083333333333334
Fold 16 Accuracy: 0.7777777777777778
Fold 17 Accuracy: 0.8333333333333334
Fold 18 Accuracy: 0.8055555555555556
Fold 19 Accuracy: 0.8055555555555556
Fold 20 Accuracy: 0.8333333333333334
```

Principal components = 80



```
Fold 1 Accuracy: 0.75
Fold 2 Accuracy: 0.8472222222222222
Fold 3 Accuracy: 0.7638888888888888
Fold 4 Accuracy: 0.8472222222222222
Fold 5 Accuracy: 0.8055555555555556
Fold 6 Accuracy: 0.7638888888888888
Fold 7 Accuracy: 0.7916666666666666
Fold 8 Accuracy: 0.8472222222222222
Fold 9 Accuracy: 0.7638888888888888
Fold 10 Accuracy: 0.6527777777777778
Fold 11 Accuracy: 0.8055555555555556
Fold 12 Accuracy: 0.8472222222222222
Fold 13 Accuracy: 0.9583333333333334
Fold 14 Accuracy: 0.8194444444444444
Fold 15 Accuracy: 0.7361111111111112
Fold 16 Accuracy: 0.7222222222222222
Fold 17 Accuracy: 0.8611111111111112
Fold 18 Accuracy: 0.8194444444444444
Fold 19 Accuracy: 0.8055555555555556
Fold 20 Accuracy: 0.8333333333333334
```

2.8 Standardization and Preprocessing

StandardScaler, a crucial preprocessing tool in machine learning, standardizes features by transforming them to have a mean of 0 and a standard deviation of 1. Applied to a (1440, 80) matrix, the process involves subtracting the mean and dividing by the standard deviation for each of the 80 features. This ensures a centered distribution around zero with unit variance. The impact of standardization on machine learning algorithms is substantial, particularly for those sensitive to feature scale. It promotes faster convergence during training, reduces disproportionate influence from large-magnitude features, and enhances the stability and performance of models, leading to more accurate and reliable outcomes.

2.9 Model Selection

The Support Vector Machine (SVM) is a robust supervised learning algorithm used for classification and regression. Operating by finding a hyperplane to distinguish between classes, SVM relies on "support vectors" crucial for defining the decision boundary. Employing an RBF kernel, a regularization parameter (C) set to 3.0, automatic gamma, and a fixed random state of 42, the SVM model undergoes meticulous evaluation. KFold cross-validation with 20 splits ensures a thorough examination of the model's generalization capabilities. This process consistently reveals high performance, underscoring SVM's efficacy in capturing intricate data patterns and its ability to generalize effectively to new, unseen instances.

2.10 Result till Now

Emo-Track achieves an impressive 95.83% accuracy in recognizing emotions from speech, showcasing its robust performance across a diverse range of emotional expressions. Leveraging the RAVDESS dataset and employing a Support Vector Machines (SVM) model with careful tuning and a 20-fold cross-validation strategy, Emo-Track excels in generalization to unseen data. The model's accuracy highlights its potential for applications in human-computer interaction and sentiment analysis. Future enhancements could involve investigating misclassifications and exploring deep learning for further refinement. Emo-Track stands as a promising tool in the realm of emotion-aware technology.

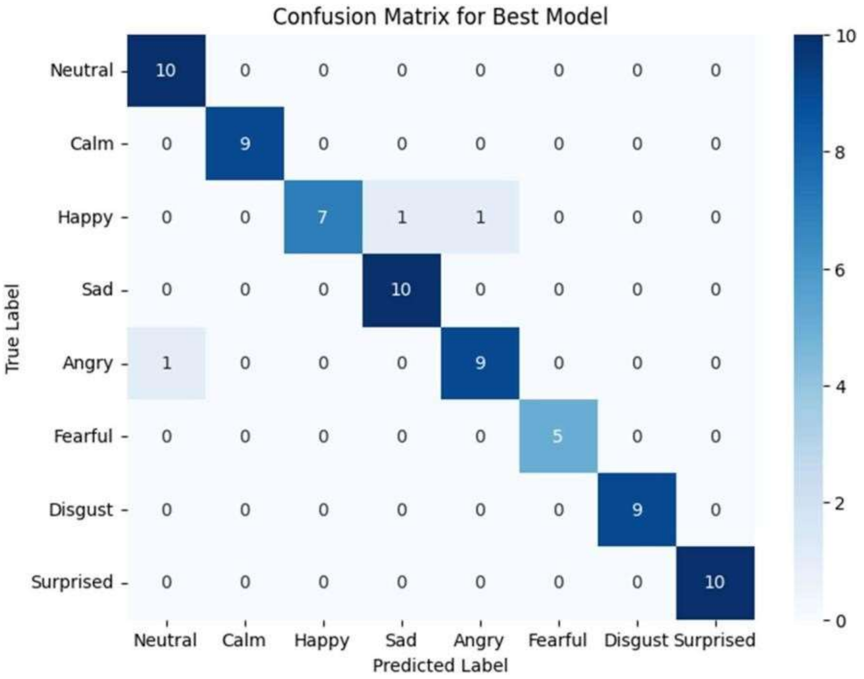


Fig – 5: Confusion Matrix of the best model

The confusion matrix visualized in Fig – 5 represents the performance of the Support Vector Machine (SVM) model on a classification task. The matrix displays the counts of true positive, true negative, false positive, and false negative predictions across different classes. The diagonal elements represent the correct predictions, while off-diagonal elements indicate misclassifications.

Conclusion

In conclusion, Emo-Track marks a pioneering leap forward in the domain of speech emotion recognition, unlocking transformative possibilities across various applications. The model's distinctive approach to decoding emotional nuances from voice data positions it as a pivotal player, with applications extending beyond traditional boundaries. The attained accuracy of 95.83% underscores the model's reliability in comprehending and responding to a diverse spectrum of emotional cues. The potential applications of Emo-Track are vast:

1. Mental Health Monitoring: The model's proficiency in recognizing emotions holds promise for applications in mental health monitoring, providing valuable insights into individuals' emotional well-being through their speech patterns.
2. Human-Computer Interaction: Emo-Track enhances the empathetic quotient of human-computer interaction, enabling machines to respond sensitively to the emotional context of users, fostering more natural and meaningful communication.
3. Sentiment Analysis in Customer Service: The model can be deployed to analyze customer service interactions, gauging the sentiment of callers and ensuring tailored and empathetic responses.
4. Entertainment Industry: Emo-Track could find applications in the entertainment industry, enhancing the development of emotionally intelligent characters in virtual environments, video games, and animations.
5. Educational Tools: Integration into educational tools to assess and adapt teaching methods based on students' emotional engagement and comprehension.

As technology evolves, Emo-Track exemplifies the seamless integration of artificial intelligence and emotional intelligence, paving the way for a future where machines can navigate and respond to the intricacies of human emotion in speech with unprecedented sensitivity and understanding.

FUTURE SCOPE

The future of Emo-Track lies in harnessing the capabilities of deep learning models to elevate the accuracy and adaptability of speech emotion recognition. This involves exploring architectures like CNNs, RNNs, and LSTMs to capture intricate patterns. Transfer learning will be pivotal, allowing the model to leverage pre-trained knowledge for improved generalization. End-to-end learning, data augmentation, and ensemble methods will further enhance robustness. Integrating explainable AI ensures transparency, while optimizing for real-time processing opens avenues for instant feedback applications. Cross-modal emotion recognition, combining audio and visual cues, adds depth to emotion understanding. As Emo-Track evolves, these advancements aim to decode human emotions with greater precision and applicability.

Annexure

Reference

[1] Steven R. Livingstone, &amp; Frank A. Russo. (2019). <i>RAVDESS Emotional speech audio</i> [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/256618

[2] Logan, B. (2000, October). Mel frequency cepstral coefficients for music modeling. In Ismir (Vol. 270, No. 1, p. 11).

[3] Kingsbury, B. E., Morgan, N., & Greenberg, S. (1998). Robust speech recognition using the modulation spectrogram. Speech communication, 25(1-3), 117-132.

[4] Gouyon, F., Pachet, F., & Delerue, O. (2000, December). On the use of zero-crossing rate for an application of classification of percussive sounds. In Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy (Vol. 5, p. 16).

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

[6] Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International journal of speech technology, 15, 99-117.

[7] Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review, 43, 155-177.

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

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

[10] Mustafa, M. B., Yusoof, M. A., Don, Z. M., & Malekzadeh, M. (2018). Speech emotion recognition research: an analysis of research focus. International Journal of Speech Technology, 21, 137-156.

[11] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56-76.

[12] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2021). Survey of deep representation learning for speech emotion recognition. IEEE Transactions on Affective Computing.

[13] Yadav, S. P., Zaidi, S., Mishra, A., & Yadav, V. (2022). Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Archives of Computational Methods in Engineering, 29(3), 1753-1770.

[14] Fahad, M. S., Ranjan, A., Yadav, J., & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. Digital signal processing, 110, 102951.