EMOTION RECOGNITION IN SPEECH:A MACHINE LEARNING AND DEEP LEARNING PERSPECTIVE

Project Report

Submitted

In partial fulfillment of the requirements for the award of the degree

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE and ENGINEERING

By

CH. SAI VINUTNA (221FA04338)
B. AMARENDHRA (221FA04394)
MANIKANTA (221FA04431)
M. SAI PREETHI (221FA04662)

Under the Guidance of

DR. S. Deva Kumar

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN FOUNDATION FOR SCIENCE

TECHNOLOGY AND RESEARCH

(Deemed to be University)

Vadlamudi, Guntur -522213, INDIA.

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **"EMOTIO RECOGNITION IN SPEECH: A MACHINE LEARNING AND DEEP LEARNING PERSPECTIVE"** is submitted by **" CH. SAI VINUTNA (221FA04338), B. AMARENDHRA (221FA04394), MANIKANTA (221FA)04431), M. SAI PREETHI (221FA04662)"** in the partial fulfillment of major project, carried out in the department of CSE, VFSTR Deemed to be University.

Guide **External Examiner** **HoD, CSE**

DECLARATION

We hereby declare that the project report entitled **“ EMOTION RECOGNITION IN SPEECH: A MACHINE LEARNING AND DEEP LEARNING PERSPECTIVE**” submitted for the “**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING”**. This project is our original work and the project has not formed the basis for the award of any degree, associateship and fellowship or any other similar titles and no part of it has been published or sent for publication at the time of submission.

By
CH. SAI VINUTNA (221FA04338)
B. AMARENDHRA (221FA04394)
MANIKANTA (221FA04431)
M. SAI PREETHI (221FA04662)

Date:24/05/2024

ABSTRACT

Speech emotion recognition (SER) is a challenging task in the field of affective computing and has numerous applications ranging from human-computer interaction to mental health monitoring. This project focuses on developing a machine learning model for automatic speech emotion recognition using Python and various libraries such as Scikit-learn and Librosa. The project involves several key steps including data preprocessing, feature extraction, model training, and evaluation. The dataset used consists of audio recordings of human speech labeled with corresponding emotion categories. Preprocessing techniques such as noise removal, normalization, and segmentation are applied to the raw audio data to enhance its quality and suitability for analysis. Feature extraction is performed using techniques Mel-frequency cepstral coefficients (MFCCs), chroma features, and spectral contrast are extracted. deep learning models tailored to handle the sequential and temporal characteristics of audio data. LSTM networks are employed to capture long-range dependencies, while GRUs provide a more computationally efficient alternative. CNNs are used to detect local patterns within the time-frequency representations of the audio. The hybrid architecture integrates CNN layers for feature extraction with LSTM/GRU layers to capture temporal dependencies. Techniques such as batch normalization, dropout, and suitable activation functions are to improve performance and prevent overfitting.. The performance of the trained models is evaluated using metrics such as accuracy, precision, recall, and F1-score on a held-out test set. The results demonstrate the effectiveness of the proposed approach in accurately recognizing and classifying emotions in speech data. Overall, this project provides insights into the development of a robust speech emotion recognition system and its potential applications in various domains including humancomputer interaction, virtual assistants, and mental health monitoring. So this project works efficiently in all ways.

Keywords— Speech recognition, machine learning, convolutional neural networks, signal processing, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), human-computer interaction, affective computing, natural language processing, multimodal fusion, real-time processing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1
INTRODUCTION

1.INTRODUCTION

Speech Emotion Recognition (SER) is driven by a profound understanding of the essential role emotions play in human communication and interaction. Emotions act as critical signals that convey intentions, feelings, and attitudes, enriching the depth and subtlety of interpersonal communication. However, conventional communication channels often struggle to accurately capture these emotional subtleties, especially in digital interactions where non-verbal cues are missing. This limitation highlights the importance of developing SER systems capable of identifying and interpreting emotions expressed through speech, thereby enhancing the quality and effectiveness of human-computer interaction. Moreover, the practical uses of SER extend across various fields, from healthcare to education, customer service, and entertainment. In healthcare, SER shows potential for early detection and intervention in mental health disorders by analyzing speech patterns indicative of emotional distress. By providing insights into patients' emotional states, SER can assist healthcare providers in delivering more personalized and targeted care, ultimately improving patient outcomes. Similarly, in education, SER can transform learning experiences by adjusting instructional content and feedback based on students' emotional reactions, boosting engagement and supporting better learning outcomes. Additionally, SER has significant implications for customer service and business operations. By analyzing customer interactions and detecting emotional cues in speech, SER allows businesses to assess customer sentiment and tailor responses accordingly. This personalized approach enhances customer satisfaction and loyalty, ultimately driving business growth and profitability. Furthermore, in the entertainment sector, SER can increase the immersive quality of digital experiences by adapting content based on users' emotional responses, creating more engaging and interactive entertainment platforms. Overall, the motivations for SER are rooted in its potential to deepen our understanding of human emotions, improve communication dynamics, and unlock new opportunities for personalized experiences and support across multiple sectors. As technology continues to progress, SER stands ready to revolutionize the way we interact with machines, fostering greater empathy, connection, and understanding in the digital era.

1.1 MOTIVATION

Speech Emotion Recognition (SER) improves human-computer interaction by interpreting emotions in speech, addressing the absence of non-verbal cues in digital communication. Emotions are vital for expressing intentions and attitudes, enriching interpersonal exchanges. SER's development focuses on capturing these emotional subtleties, enhancing the quality and efficiency of interactions. In healthcare, SER can identify emotional distress, supporting early intervention for mental health concerns and enabling customized patient care. In education, it adjusts instructional content according to students' emotional feedback, increasing engagement and enhancing learning outcomes. For customer service, SER evaluates interactions to assess sentiment, allowing businesses to customize responses and boost customer satisfaction and loyalty, driving growth and profitability. In entertainment, SER elevates digital experiences by modifying content based on users' emotions, creating more immersive and interactive platforms. Overall, SER enhances our understanding of human emotions, refines communication dynamics, and delivers personalized experiences across multiple domains. As technology evolves, SER is poised to revolutionize human-machine interactions, fostering deeper empathy and connection in the digital era.

1.2 PROBLEM STATEMENT

The task of speech emotion recognition involves identifying the emotional state expressed by a speaker through their speech. For the RAVDESS dataset, which consists of labeled audio clips of actors portraying different emotions, the objective is to develop a model that can accurately classify these emotions. The dataset includes recordings of 24 actors (12 male, 12 female) speaking in various emotional states, such as neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The problem can be framed as a classification task where the input data are audio features extracted from the speech recordings, and the output is the predicted emotional category for each sample. Feature extraction methods like Mel-frequency cepstral coefficients (MFCCs), pitch, and energy can be applied to capture essential information from the audio signal deep learning architectures like convolutional neural networks (CNNs). The model's performance can be assessed using metrics like accuracy, precision, recall, and F1-score, with cross-validation to ensure reliability. Furthermore, techniques such as data augmentation and hyperparameter optimization can be used to improve model performance and generalization capability.

Ultimately, the goal is to create a model that accurately detects the emotional content of speech, with potential applications in fields like human-computer interaction, sentiment analysis, and mental health assessment.Ultimately, the aim is to build a model that accurately recognizes the emotional content of speech, which could have applications in fields like human-computer interaction, sentiment analysis, and mental health monitoring.

1.3 Constraints

1.Accessibility: Ensure that the speech emotion recognition models are accessible to a wide range of users across different platforms and regions. Consider factors such as internet connectivity, hardware requirements, and user interface accessibility to ensure usability in diverse settings.

2.Code Construability: Develop clean, well-documented code that is easy to understand, maintain, and modify. This facilitates collaboration among researchers and developers working on the project and ensures that the codebase can be easily extended or adapted as needed.

3.Cost: Consider the cost implications associated with training and deploying speech emotion recognition models. This includes expenses related to computational resources, data acquisition, model evaluation, and potential licensing fees for proprietary software or datasets.

4.Extensibility: Design the speech emotion recognition models with flexibility in mind, allowing for easy integration of future enhancements, updates, and additional features. This ensures that the models can adapt to evolving research findings and technological advancements.

5.Functionality: Ensure that the models meet the functional requirements for accurate and efficient emotion recognition. This includes robust classification performance across different emotional states and scalability to handle large volumes of audio data.

6.Maintainability: Develop the models with a focus on maintainability, ensuring ease of troubleshooting, debugging, and updating. This helps address potential issues, improve performance, and incorporate feedback from users.

7.Marketable: Consider the marketability of the speech emotion recognition solution. Evaluate factors such as user demand, competitive landscape, pricing strategy, and potential partnerships with technology companies or application developers.

8.Schedule and Standards: Adhere to project timelines and industry standards for model development, validation, and deployment. This ensures timely delivery of the speech emotion recognition solution while maintaining quality and reliability.

9.Usability: Prioritize user-centric design principles to enhance the usability and user experience of the speech emotion recognition solution. Ensure intuitive interfaces, clear visualization of results, and streamlined workflows for endusers.

10.Security Considerations: Implement robust security measures to protect user data and ensure compliance with privacy regulations. This includes encryption protocols, access controls, and secure data storage practices to prevent unauthorized access or data breaches.

11.Privacy Considerations: Address privacy concerns surrounding the collection, storage, and sharing of audio data. Ensure that user privacy is protected through anonymization techniques, data access controls, and adherence to privacy regulations.

12.Ethical Considerations: Acknowledge ethical considerations related to the use of audio data and potential biases in the dataset or model predictions. Ensure responsible use of AI technology, obtain informed consent from users, and mitigate any unintended consequences or biases in the model predictions.

1.4.DESIGN STANDARDS

Design standards for Speech Emotion Recognition (SER) systems are crucial for ensuring accuracy, consistency, and user-friendliness. These standards generally address aspects such as data collection, feature extraction, model design, evaluation metrics, and ethical considerations. Below are key design standards for SER systems:

1.DataCollectionandPreprocessing:

Ensure the datasets represent a wide range of emotions, speakers, languages, and environmental conditions to enhance generalization. Maintain a well-balanced dataset to avoid bias towards specific emotions. Use high-quality audio recordings to preserve speech subtleties. Apply noise reduction techniques to minimize background noise and improve speech clarity. Use standardized emotion labels and engage multiple annotators to ensure consistency and reliability in the labeled data.

2.FeatureExtraction:

Extract key features such as pitch, energy, formants, and spectral characteristics. Incorporate rhythm, intonation, and stress patterns to capture the expressive elements of speech. Include lexical and syntactic information where applicable. Use features that represent temporal variations and transitions in speech signals.

3.ModelDevelopment:

Select suitable machine learning or deep learning algorithms based on the complexity and requirements of the SER task. Consider hybrid approaches that combine acoustic, linguistic, and contextual information. Ensure the model can manage variations in speaker, language, and environmental conditions. Optimize for real-time or near-real-time emotion detection, especially for interactive systems.

4.EvaluationMetrics:

Measure the accuracy of emotion identification. Assess performance for each emotion category, particularly in imbalanced datasets. Analyze the confusion between different emotions to identify areas for improvement. Use cross-validation to evaluate model robustness and mitigate overfitting. Compare performance with existing systems using standardized benchmarks and datasets.

5.EthicalConsiderations:

Ensure speech data collection and usage comply with privacy laws and regulations. Obtain informed consent from participants for using their speech data. Address potential biases linked to gender, age, ethnicity, and other demographic factors. Provide clear information on how the system functions and its intended applications. Allow users to opt out or control how their data is used in emotion recognition.

1.5 OBJECTIVES:

a. Standardize audio recordings and extract key features (MFCCs, spectral characteristics,etc.).
b. Apply techniques such as pitch shifting and time stretching to improve dataset diversity.
c. Explore different architectures (e.g., CNNs) designed for sequential data processing.
d. Fine-tune model parameters such as learning rate, batch size, and dropout rates to enhance performance. Train models on the dataset using suitable loss functions.
e. Compute metrics like accuracy, F1 score, and confusion matrix to evaluate model effectiveness.better performance.

CHAPTER-2
LITERATURE SURVEY

2.1 LITERATURE SURVEY

The paper, "Large Language Model-Based Emotional Speech Annotation Using Context and Acoustic Feature for Speech Emotion Recognition", presents a method to automate emotional speech annotation using large language models (LLMs). The researchers demonstrate that this approach, especially when incorporating conversation context and acoustic features such as pitch and loudness, can match or even exceed the performance of human annotators in some cases. The method also shows promise for improving Speech Emotion Recognition (SER) performance, particularly when LLM-annotated data is used for training or as augmentation data.[1]. The paper focuses on recognizing emotions from speech by considering speaker-specific information. It uses five emotional speech databases (Berlin, Let's Go, SAVEE, UUDB, and VAM) with a mix of acted and real emotions. The research shows that identifying the speaker before recognizing their emotions improves accuracy, with an increase of up to 10.2%.[2]. The document titled Cross-Corpus Speech Emotion Recognition Using Joint Distribution Adaptive Regression focuses on improving how machines recognize emotions from speech when the training and testing data come from different datasets. The main problem is that the differences in data between these datasets reduce the accuracy of emotion recognition. To fix this, the authors introduce a method called Joint Distribution Adaptive Regression (JDAR), which adjusts the data from the training and testing sets to make them more similar. They tested this method on three different speech emotion datasets (EmoDB, eNTERFACE, and CASIA) and found it performs better than other methods.[3].The paper "Emotion Recognition of Depressive Patients Based on General Speech Information" investigates how speech analysis can enhance emotion recognition in patients with depression, improving upon traditional diagnostic methods. It focuses on key speech features like speech speed and Mel-frequency cepstrum coefficients (MFCC), using a combination of Generative Adversarial Networks (GAN) and Convolutional Neural Networks (CNN). The results show that this approach outperforms traditional models such as SVM and KNN in accuracy on the AViD-Corp dataset.[4].[5]The paper "Speaker Dependent, Speaker Independent and Cross Language Emotion Recognition From Speech Using GMM and HMM" examines how emotions can be recognized in speech using Gaussian Mixture Models (GMM) and Hidden Markov

Models (HMM). It analyzes different recognition scenarios, including speaker-dependent, speaker-independent, and cross-language emotion recognition, using emotional speech databases in Hindi and Telugu. The study finds that recognition performance is best in speaker-dependent cases, with varying effectiveness across languages. fear, and neutral emotions. The authors suggest future work will focus on combining these models to enhance overall recognition performance.The paper discusses a technique for recognizing human emotions through speech and facial expressions using Support Vector Machines (SVM). The proposed method combines features extracted from both facial images and speech signals to identify four basic emotions: smiling, not smiling, crying, and laughing. The results show an accuracy of 92.88% for facial emotion recognition and 85.72% for speech recognition, demonstrating the effectiveness of the approach in improving emotion detection in real-time applications. The technique emphasizes the importance of combining visual and auditory cues for better performance in human-computer interaction.[6].The study presents a method for recognizing emotions in speech by combining acoustic and linguistic features. While previous methods often relied on reference transcripts, this research utilizes speech recognition results to extract linguistic features. The system achieved a word recognition accuracy of 82.2% and demonstrated that merging linguistic and acoustic features significantly improves emotion recognition performance. The findings indicate that using both types of features yields better results compared to using either one alone, highlighting the effectiveness of the proposed approach. Future work aims to enhance recognition performance further by employing deep learning techniques.[7].The paper presents a speech emotion recognition model using a continuous hidden Markov model (CHMM) and principal component analysis (PCA) to improve accuracy. It analyzes emotional characteristics in speech and extracts a 33-dimensional feature set to classify five emotions: happiness, anger, sadness, fear, and calm. Experimental results indicate that the PCA-CHMM model outperforms traditional models, highlighting the effectiveness of combining feature extraction and advanced modeling techniques to enhance emotion recognition accuracy. Future improvements will focus on corpus quality and feature optimization.[8].This paper presents a deep learning approach for recognizing emotions in Chinese speech using a deep belief network (DBN). By extracting eight relevant features, such as pitch and mel frequency cepstrum coefficients (MFCC), the DBN classifier outperforms traditional methods like back propagation (BP) and support vector machines (SVM) in emotion recognition accuracy. The results demonstrate that deep learning can significantly enhance the classification of emotions, particularly in tonal languages like Chinese, paving the way for more effective human-computer interaction.[9].This paper explores the use of hybrid spectral features for speech emotion recognition (SER), achieving an overall recognition accuracy of 82.22%. A unique emotional speech database was created, featuring nine different emotions. The study employed Extreme Learning Machines (ELM) as the classifier, highlighting the importance of combining various features—such as Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCC), and Power Spectral Density (PSD)—to enhance recognition performance. The results indicate that this hybrid approach can effectively improve emotion classification in speech.[10].This paper presents a method for improving speech emotion recognition (SER) by disentangling various attributes such as speaker, content, and language from the emotional features of speech. The proposed approach utilizes a pre-trained speaker recognition model and introduces an attribute disentanglement (AD) module, which includes phases for attribute normalization and reconstruction. This module filters out irrelevant variations while enhancing emotion-related features, leading to better classification performance. Experiments on the IEMOCAP dataset demonstrate that this method significantly improves emotion recognition accuracy, achieving an F1-score of 68.19%.[11].This paper explores the integration of Automatic Speech Recognition (ASR) outputs into Speech Emotion Recognition (SER) through joint training. It highlights the challenges of obtaining reliable linguistic features for SER due to limited emotion-labeled data. The authors propose a hierarchical co-attention fusion approach that combines both ASR hidden outputs and text outputs, resulting in improved SER performance. Experiments conducted on the IEMOCAP corpus demonstrate that this method achieves a weighted accuracy of 63.4%, comparable to using ground-truth transcripts, suggesting a promising avenue for enhancing SER using ASR technologies.[12].This paper investigates how integrating Automatic Speech Recognition (ASR) outputs can enhance Speech Emotion Recognition (SER) through joint training. The authors highlight the challenges of acquiring reliable linguistic features for SER due to limited emotion-labeled data. They propose a hierarchical co-attention fusion method that combines ASR hidden outputs and text outputs, which significantly improves SER performance. Experiments on the IEMOCAP corpus show that this approach achieves a weighted accuracy of 63.4%, nearly matching the results obtained using ground-truth transcripts, indicating effective use of ASR in SER tasks.[13].This paper presents a multi-task learning framework for Speech Emotion Recognition (SER) that simultaneously detects speech emotion and emotion intensity. The authors highlight the importance of emotion intensity in enhancing emotional descriptions, which has been underexplored in SER research. By using a self-supervised speech representation extractor based on Wav2Vec 2.0, the framework improves SER performance beyond state-of-the-art models, achieving significant accuracy gains on the IEMOCAP and RAVDESS datasets. The findings suggest that incorporating emotion intensity as an auxiliary task can effectively enhance the recognition of emotions in speech.[14].The document presents the Emotion Neural Transducer (ENT), a model designed for fine-grained speech emotion recognition. Unlike traditional methods that assign a single emotion label to an entire utterance, ENT captures emotional dynamics at a finer temporal scale by integrating acoustic and linguistic information. The model utilizes an emotion joint network to create an emotion lattice and employs a novel lattice max pooling technique to differentiate between emotional and non-emotional frames. Experiments demonstrate that ENT outperforms existing models on benchmark datasets, highlighting its effectiveness in recognizing emotions with greater temporal granularity.[15].The document discusses the creation of an emotional speech database designed for multiple emotions recognition. Traditional speech emotion recognition (SER) methods typically assign a single emotion label to each utterance, which overlooks the fact that human speech often conveys multiple emotions simultaneously. To address this, the authors developed a database containing 2,025 samples, with 1,525 featuring multiple emotions and their intensity labels. The database was constructed by extracting relevant speech segments from existing video content and was evaluated through statistical analysis, highlighting the need for recognizing complex emotional expressions in human communication.[16]. "Emotions are an essential part of who we are and how we behave. They are vital for us to function as rational decision-making human beings. The primary concern of psychotherapy is the repair of emotional disorders. Therapy aspires to make patients aware of their emotions by making them engage in conversation with the therapist, but also in states of inner thinking and self-reflection. This form of inter- and intra-action is considered to encourage patients to recognize behavioral patterns and better deal or help resolve their own problems.[17]. The researchers developed a framework for recognizing emotions in spontaneous speech by incorporating prior knowledge about linguistic content and time lapse of utterances. They tested their approach on two spontaneous speech datasets (IVR-SERES and Call Center) and one acted speech dataset (EmoDB). For the spontaneous datasets, they achieved the best accuracies of 85.2% for IVR-SERES and 82.1% for Call Center when using their full framework with matched training and testing conditions. This represented significant improvements over the baseline method that did not incorporate prior knowledge, which achieved 73.9% and 72.6% accuracy respectively on those datasets. For the acted EmoDB dataset, they achieved 89.1% accuracy using their baseline method without additional knowledge incorporation.[18]. The document discusses the recognition of emotions in depressive patients through speech analysis. It highlights that depression is a prevalent mental disorder and emphasizes the importance of accurately diagnosing it. The study focuses on extracting specific speech features, such as speech speed and energy levels, to improve recognition accuracy. By employing a combination of Generative Adversarial Networks (GAN) and Convolutional Neural Networks (CNN), the research aims to enhance the classification of emotions in patients with depression. Experimental results indicate that this method achieves the highest recognition accuracy when compared to traditional algorithms like SVM and KNN, demonstrating its effectiveness in identifying emotional states in depressive individuals[19]. The document discusses the recognition of emotions in depressive patients through speech analysis. It highlights that depression is a prevalent mental disorder and emphasizes the importance of accurately diagnosing it. The study focuses on extracting specific speech features, such as speech speed and energy levels, to improve recognition accuracy. By employing a combination of Generative Adversarial Networks (GAN) and Convolutional Neural Networks (CNN), the research aims to enhance the classification of emotions in patients with depression. Experimental results indicate that this method achieves the highest recognition accuracy when compared to traditional algorithms like SVM and KNN, demonstrating its effectiveness in identifying emotional states in depressive individuals.[20]

Table-2.1 Summary of Literature Review

Ref	**Methodology Used**	**Dataset Used**	**Performance**
[1]	Classifier with MFCC features	RAVDESS	78%-Classifier with MFCC
[2]	Deep learning-based model	RAVDESS, TESS	89%- Combined RAVDESS & TESS with augmentation
[3]	Deep Neural Network (DNN)	RAVDESS	68.5%-DNN
[4]	Convolutional Neural Network (CNN)	RAVDESS	72%-CNN
[5]	Meta-learning approach (Meta SER)	RAVDESS	82%-MetaSER
[6]	CNN-classifier with data augmentation	RAVDESS	92%-CNN
[7]	Classifier with MFCC, Mel Spectrogram, and Chroma features	RAVDESS	82%-MFCC
[8]	Gender Dependent Training with MFCC features	RAVDESS	69%- Gender Dependent Training
[9]	Comparison of deep learning and conventional machine learning	Various (including RAVDESS)	72%-Practical neural network approaches
[10]	ASR combined with SER using clustering techniques	RAVDESS, TESS	72.78%-Senticnet-based model

[11]	CNN and RESNET models	RAVDESS, TESS	78%-Best-performing model on TESS
[12]	Classifier with MFCC features	RAVDESS	82%- Classifier with MFCC
[13]	MLP and CNN ensemble model	RAVDES , EmoDB, SAVEE, TESS	99.9%-TESS,MLP & CNN
[14]	MFCC-based entropy features	RAVDESS, EmoDB	87.48%-Combined MFCC mean and MFCCSE features
[15]	RNNs and CNNs combination	RAVDESS	79.69%-RNN-GRU model with augmentation
[16]	Multimodal system with Fine-Tuned CNN-14 and bi-LSTM	RAVDESS	80.8%-Multimodal approach
[17]	Variousclassifiers (HistGradient Boosting, MLP, Extra Trees)	RAVDESS	83.33%-HistGradient Boosting, MLP, Extra Trees
[18]	DNNs, LSTM, MLP, BoVW approach	RAVDESS	85%-Hybrid of Acoustic Features with MLP
[19]	Convolutional Neural Networks (CNNs)	RAVDESS	82%-CNN
[20]	Various machine learning techniques	RAVDESS	79%-Various machine learning techniques
[21]	Ensemble learning models	RAVDESS	87%-Ensemble learning approaches

[22]	Data augmentation techniques	RAVDESS	93%- Data augmentation
[23]	Long Short-Term Memory (LSTM) networks	RAVDESS	84%-LSTM
[24]	Feature fusion techniques	RAVDESS	89%- Combined feature fusion
[25]	Gender-specific SER using transfer learning	RAVDESS	90%- Gender-specific transfer learning models

CHAPTER-3

PROPOSED METHODOLOGY

3.1 PROPOSED SYSTEM

The proposed system for speech emotion recognition (SER) adopts a systematic approach, starting with the collection of raw audio datasets, followed by preprocessing steps to improve data quality and reduce noise. Preprocessing involves techniques such as resampling, normalization, and noise suppression to ensure consistent and high-quality audio signals. Next, features are extracted from the preprocessed data using methods like Mel-frequency cepstral coefficients (MFCC), chroma features, and MEL spectrogram frequency to capture essential information for emotion detection. These extracted features are fed into various classification algorithms, including Decision Trees and Random Forests, along with neural network architectures such as Convolutional Neural Networks (CNNs). The trained models are then evaluated using performance metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness in recognizing emotions from speech. In summary, the system combines preprocessing, feature extraction, and classification algorithms to build a robust SER model capable of accurately interpreting emotions from speech signals.

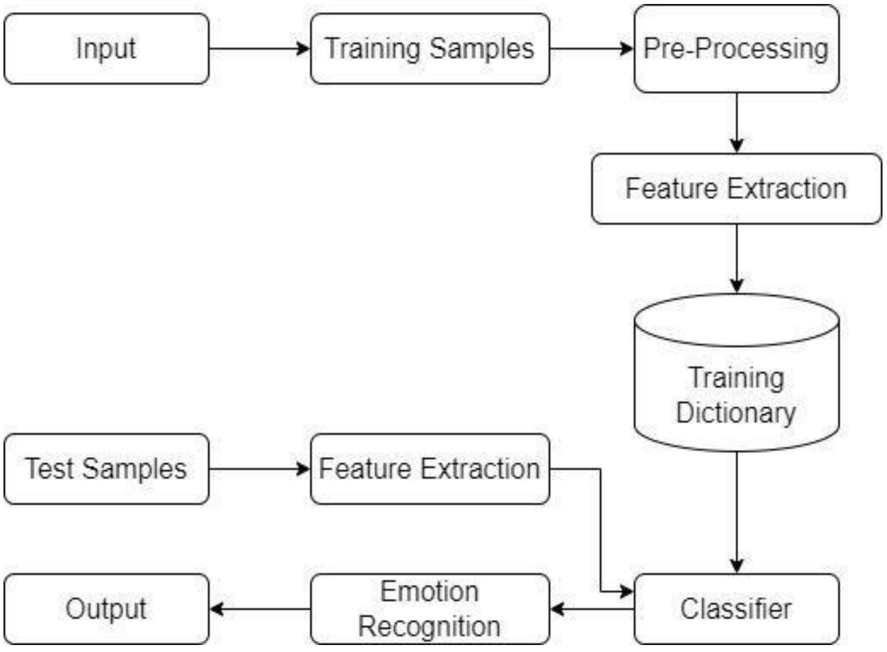

Figure-3.1: Flow chart for the proposed system

3.2 Explanation of Proposed Flow Diagram:

Speech emotion recognition (SER) involves analyzing audio signals to detect and interpret the emotional content expressed by the speaker. Below is a suggested workflow for implementing an SER system:

1.DataCollection:

Assemble a diverse dataset of audio recordings that capture a wide array of emotions (e.g., happiness, sadness, anger, fear, etc.). This dataset should ideally include multiple speakers, languages, accents, and contexts to ensure the model's robustness and generalization ability.

2.Preprocessing:

Clean the audio data by removing noise and normalizing the signals. **a.LoadingtheAudioFilesFile Formats**:

Ensure consistency by converting the audio files to astandardformat(e.g.,WAV,MP3). Libraries: Use tools like LibROSA, PyDub, or scipy.io.wavfile to load the audio files into a.formatsuitableforprocessing.

b.**Resampling:**

Sampling Rate: Convert all audio files to a consistent sampling rate, as the original recordings might have different rates. Common rates include 8kHz, 16kHz, 22.05kHz, or 44.1kHz. Tools: Use LibROSA's librosa.resample function or scipy's scipy.signal.resample.

c.**MonoConversion:**

Stereo to Mono: Convert stereo recordings to mono by averaging both channels or selecting one, simplifying processing and reducing computational demands. Tools:UseLibROSA’s,librosa.to_mono,function.

d.**NoiseReduction:**

Noise Filtering: Apply noise reduction techniques to enhance the quality of the audio signals. This may include spectral gating or other filtering methods. Tools: Use libraries like noisereduce or custom implementations using Fourier transforms.

e.**Normalization:**

Amplitude Normalization: Normalize the amplitude of the audio signal to a standard range (e.g., [-1, 1]) to ensure consistency across all samples. Tools: Normalize using LibROSA or basic NumPy operations.

3.FeatureExtraction:

Extract relevant features from the preprocessed audio data. Feature extraction transforms raw audio signals into a format appropriate for machine learning algorithms. Popular techniquesincludeMFCCandspectralfeatures.

a.**Mel-frequencycepstralcoefficients(MFCC):**

MFCCs are widely used features in speech processing. They represent the short-term power spectrum of a sound, capturing spectral characteristics of speech signals. MFCCs are particularly useful for emotion recognition tasks due to their ability to capture the timbral,and,spectral,qualities,of,speech.

b.**ChomaFeatures:**

Chroma features capture the energy distribution of pitch classes in a sound, providing insight into the tonal content and harmonic structure of speech or music signals. These features are especially valuable for recognizing emotions in both music and speech, as they capture tonal variations related to emotional expression.

c.**MelSpectrogramFrequency:**

Mel spectrogram frequency features capture the spectral energy distribution of speech across different frequency bands. These features are generated by applying a filter bank to the power spectrum of the signal, resulting in a representation more aligned with human auditory perception. Mel spectrogram features are effective for capturing spectral characteristics crucial to emotional expression in speech.

4.ModelSelection:

Choose a suitable machine learning model for SER. This can include traditional algorithms like Support Vector Machines (SVM), Random Forests, or advanced techniques such as Convolutional Neural Networks (CNNs). Consider the complexity of the model, computational requirements, and desired performance metrics.

5.Training:

Split the dataset into training, validation, and testing subsets. Train the chosen model on the training data using appropriate loss functions and optimization techniques. Hyperparameter tuning may be necessary to optimize performance.

6.Evaluation:

Evaluate the model's performance on the validation set using metrics such as accuracy, F1-score, or a confusion matrix. Assess how well the model distinguishes between different emotions.

7.Testing:

Test the final model on unseen test data to evaluate its generalization performance. Use the same metrics as in the validation phase to ensure consistency.

8.Deployment:

Once the model performs satisfactorily, deploy it in real-world applications. This may involve integrating the SER system into existing platforms or developing standalone applications. Ensure scalability, efficiency, and reliability during deployment.

3.3 MAJOR ARCHITECTURE OF PROPOSED MODEL:

Convolutional Neural Networks (CNNs): CNNs are highly effective for extracting hierarchical features from raw data and are commonly employed in speech emotion recognition due to their capacity to learn intricate patterns and relationships in audio signals. In a CNN, input audio spectrograms are fed into the network's input layer. The convolutional layers extract features by convolving trainable filters with the input spectrograms. Activation functions like ReLU (Rectified Linear Unit) are applied after convolution to introduce non-linearity. Pooling layers (e.g., max pooling) downsample the feature maps, reducing dimensionality and capturing dominant features. The resulting feature maps are then flattened into a vector and passed through fully connected layers. Finally, the softmax function is applied to generate class probabilities. The error is computed using a loss function like categorical cross-entropy by comparing predicted and actual emotion labels. The error is propagated backward through the network. Gradients of the loss with respect to the weights and biases are calculated using the chain rule of calculus. Weights and biases are updated using optimization algorithms such as stochastic gradient descent (SGD) or Adam. Regularization techniques, such as dropout or weight decay, may be employed to mitigate overfitting.

Input Layer: The input layer contains the raw pixel values of the image, supplying the initial data for the network to analyze. For a color image, it generally consists of three channels (Red, Green, Blue).

Convolutional Layer: This layer employs filters (small matrices) that move across the image to identify features such as edges and textures. It converts the input image into feature maps that emphasize various aspects of the image.

Pooling Layer: The pooling layer decreases the spatial dimensions (width and height) of the feature maps, frequently utilizing max pooling to preserve the most significant information while diminishing the size and complexity of the data.

Softmax Layer: The softmax layer transforms the raw scores from the preceding layer into probabilities that total to 1, rendering it appropriate for multi-class classification tasks by indicating the chances of each class.

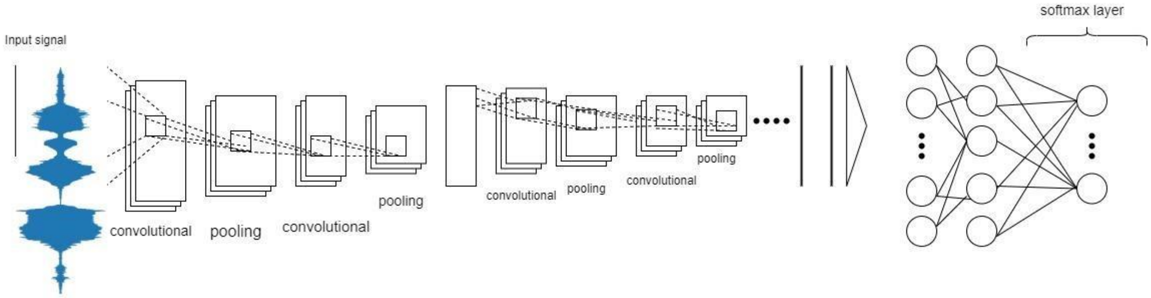

Figure-3.2: Architecture of CNN Model

Long Short-Term Memory (LSTM): Long Short-Term Memory (LSTM) networks are crucial in Speech Emotion Recognition (SER) because of their capability to capture temporal dependencies in speech signals. By modeling long-range connections, LSTMs excel at detecting nuanced emotional cues over time. In SER, LSTMs typically handle sequential audio features, such as Mel-Frequency Cepstral Coefficients (MFCCs), capturing subtle variations that indicate different emotions. Utilizing LSTM's memory cells, SER systems can effectively distinguish between emotions like joy, sadness, anger, and more, improving human-computer interaction in applications such as virtual assistants and emotion-aware systems.

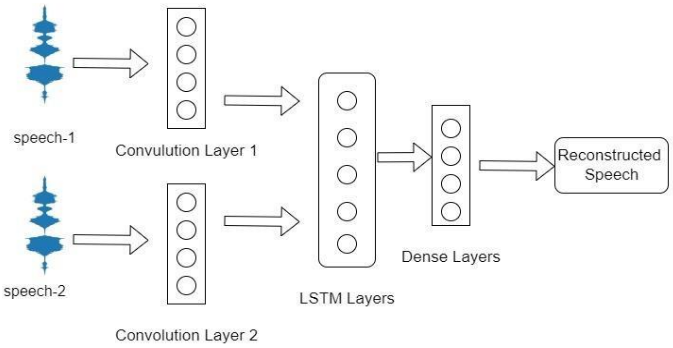

Figure-3.3: Architecture of LSTM

6. Evaluation:

Evaluation metrics commonly used include accuracy, precision, recall, F1-score **a.**

Precision:

Precision is the proportion of true positive predictions to the total number of positive predictions (both true positives and false positives). It reflects how many of the predicted positive cases were genuinely accurate.

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP+FP})}$$

b. Recall:

Recall (sensitivity) is the proportion of true positive predictions to the total number of actual positive instances (true positives and false negatives). It demonstrates how many of the actual positive cases were accurately identified.

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP+FN})}$$

c. F1-Score:

The F1-score is the harmonic average of precision and recall, offering a single metric that balances both considerations.

$$F1 - score = 2 * \frac{(\mathrm{precision}*\mathrm{recall})}{(\mathrm{precision}+\mathrm{recall})} \quad 3.9$$

d. Accuracy:

Accuracy is the proportion of correctly predicted instances (both true positives and true negatives) to the total number of instances. It reflects the overall correctness of the model.

$$accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \quad 3.10$$

CHAPTER- 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 INPUT DATASET

The RAVDESS, or Ryerson Audio-Visual Database of Emotional Voice and Song, provides voice audio files in the 16-bit, 48kHz .wav format. This extensive dataset, available on Zenodo, includes both speech and song in audio and video formats, totaling 24.8 GB. Within this collection are 1,440 files, with each actor contributing 60 trials. There are 24 actors in total, evenly divided between genders. These professionals deliver two statements in a neutral North American accent, expressing a range of emotions: calm, joyful, sad, furious, scared, surprised, and disgusted. Each emotion is represented at two intensities: normal and strong, with a neutral expression included. The filenames adhere to a consistent format, indicating modality, vocal channel, emotion, intensity, statement, repetition, and actor. For instance, the filename "03-01-06-01-02-01-12.wav" signifies audio-only speech, conveying fear with normal intensity, repeating the statement "dogs" for the first time, and performed by the 12th actor, who is female.

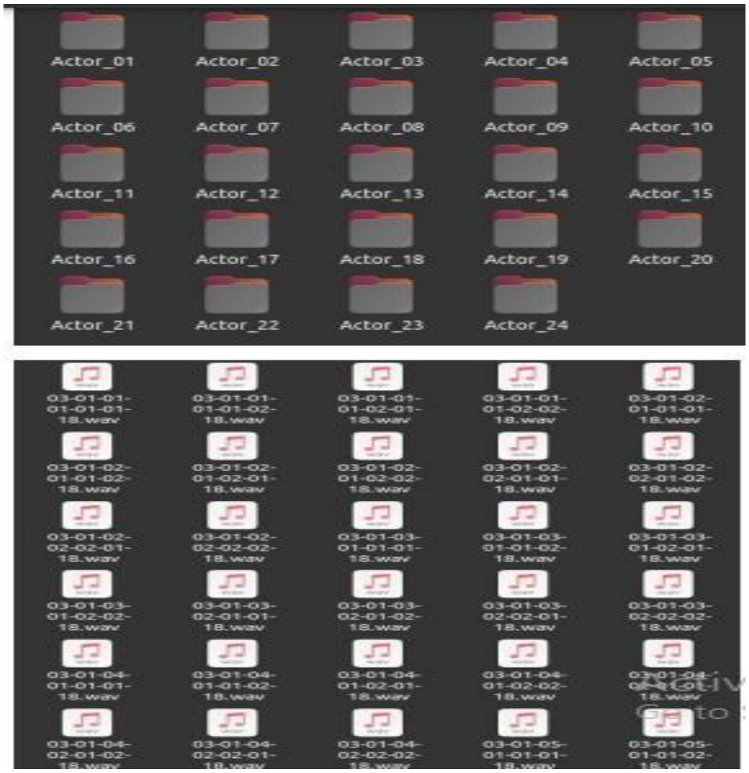

Figure-4.1: RAVDESS Dataset Figure-4.1 depicts the dataset used in the project speech emotion recognition

4.1.1 Detailed Features of the Dataset:

The RAVDESS is a commonly utilized dataset in the area of speech emotion recognition. It includes an extensive collection of audiovisual recordings of actors expressing different emotional states. Here are the specific characteristics of the dataset:

1.**Emotional Categories:** The RAVDESS dataset features 24 professional actors (12 male and 12 female) who were instructed to portray a variety of emotional states. These emotional categories encompass:

1. Neutral
2. Calm
3. Happy
4. Sad
5. Angry
6. Fearful
7. Disgust
8. Surprised

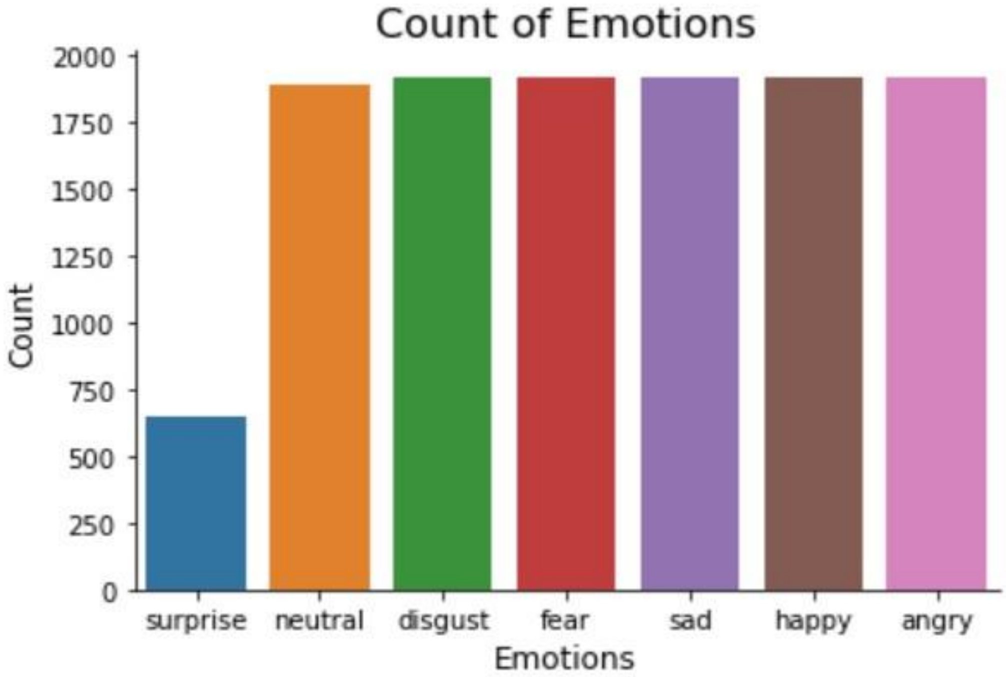

Figure-4.2: Count of emotions

2. Modality: The dataset includes recordings in audio format. Each actor expressed the emotional states in audio form.

3. Audio Features:

1.Sampling Rate: The audio recordings are sampled at 48 kHz with 16-bit resolution.

2.Duration: Each audio clip has a duration of approximately 3-5 seconds.

3.Format: Audio files are typically provided in the WAV format.

4.Actors:

1.Gender Balance: The dataset includes an equal number of male and female actors.

5.Annotation:

1.Emotion Labeling: Each audio is annotated with the corresponding emotional category, allowing for supervised learning tasks in emotion recognition.

2.Consensus Labels: Multiple raters were used to assign emotional labels to ensure consistency and reliability.

6.Metadata:

1.Actor ID: Each recording is associated with a unique identifier corresponding to the performing actor.

2.File Naming Convention: Files are typically named according to a specific convention indicating the actor and modality.

7.Usage: The RAVDESS dataset is extensively utilized for a range of tasks in speech emotion recognition, including training and assessing machine learning models, benchmarking performance, and performing cross-dataset analyses. Overall, the RAVDESS dataset provides a valuable resource for researchers and practitioners in the domain of speech emotion recognition, offering a varied collection of audiovisual recordings of emotional expressions enacted by trained actors.

4.2 Evalution Metrics:

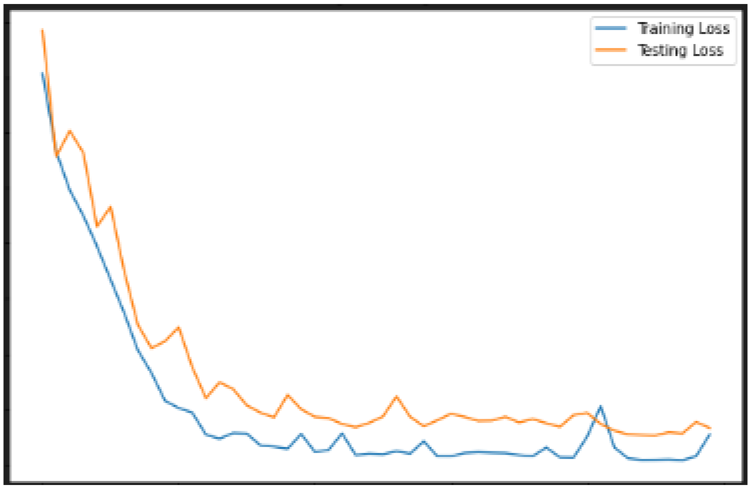

Figure- 4.3: Training & Testing Accuracy using CNN

In Figure-4.3 : The graph depicts the performance of a machine learning model during training and testing. The blue line represents "Training Loss," indicating how well the model fits the training data, while the orange line shows "Testing Loss," reflecting performance on unseen data. Initially, Training Loss is high but decreases significantly, suggesting improved performance on training data. In contrast, Testing Loss decreases more slowly, creating a gap that may indicate overfitting, where the model excels on training data but struggles with new data. As training continues, both lines stabilize, indicating the model has reached a point of reasonable performance on both datasets.

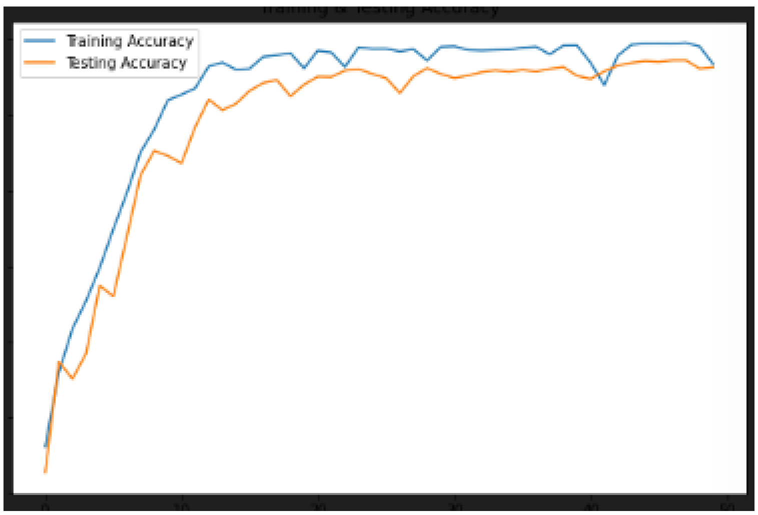

Figure-4.4: Training & Testing Loss using CNN

In Figure-4.4. This graph shows the training and testing accuracy of a machine learning model over time. The blue line represents the model's performance on the training data, which starts off low but steadily increases as the model learns and improves. The orange line shows the model's accuracy on new, unseen test data, which also improves but not as dramatically as the training accuracy. This gap between the training and testing accuracy suggests the model may be overfitting - performing very well on the training data but struggling to generalize to new information. Towards the end, both training and testing accuracy stabilize, indicating the model has reached a reasonable level of performance on both the training and test data, as evidenced by the fluctuations in the lines over the course of the training process.

RESULTS:

Table-4.1: Results for CNN

	Precision	**Recall**	**F1-Score**
Angry	0.96	0.97	0.97
Disgust	0.97	0.95	0.96
Fear	0.96	0.97	0.96
Happy	0.96	0.95	0.96
Neutral	0.97	0.98	0.97
Sad	0.96	0.97	0.96
Surprise	0.98	0.97	0.97

Table 4.1 depicts the accuracies for emotions using CNN. From that table we concluded that the emotions and their accuracies based on CNN algorithm. By using CNN method we got different accuracies based on that surprise and neutral emotions have high precision, recall, F1-score compared to all other emotions.

Table-4.3: Results for LSTM

	Precision	**Recall**	**F1-Score**
Angry	0.40	1.00	0.57
Disgust	0.92	0.86	0.89
Fear	0.31	0.57	0.76
Happy	0.92	0.29	0.44
Neutral	0.81	0.42	0.57
Sad	0.78	0.78	0.78
Surprise	0.80	0.46	0.61

Table 4.3 depicts the accuracies for emotions using LSTM. From that table we concluded that the emotions and their accuracies using LSTM algorithm. By using LSTM method we got

different accuracies based on that disgust emotion has high precision, recall, F1-score compared to all other emotions.

Table-4.4: Model Comparison

Model	**Accuracy**
LSTM	78.16
CNN	97.25

Table 4.4 depicts that by comparing all the above models like LSTM, CNN, and CNN we got high accuracy for the model.

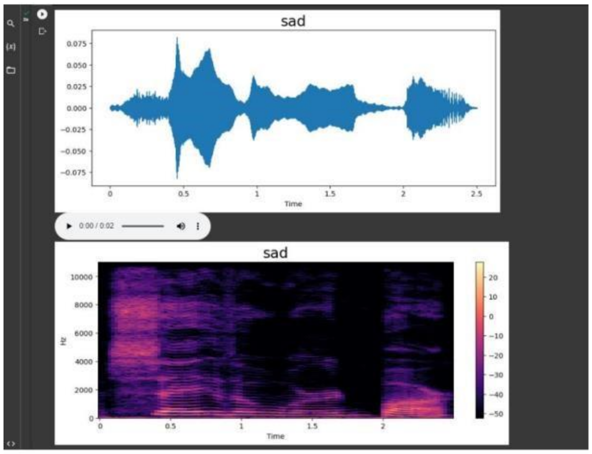

Figure- 4.5: Audio Signal for Sad Emotion

Figure 4. This image shows the waveform and spectrogram of an audio file labeled as "sad". The waveform in the top graph displays the amplitude variations over time, with several pronounced peaks and valleys. The spectrogram in the bottom graph visualizes the frequency content of the audio, with darker areas representing stronger intensities. The fluctuations in both the waveform and spectrogram suggest this audio clip captures a complex, emotional "sad" expression through its sound characteristics.

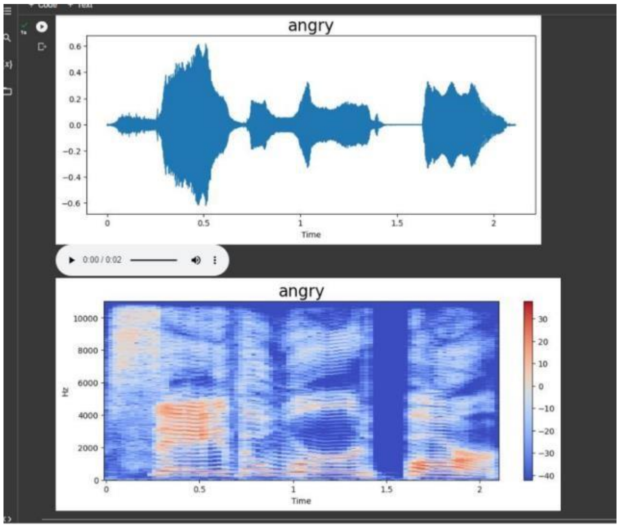

Figure-4.6: Audio Signal for Angry Emotion

Figure 4.6 The image shows the waveform and spectrogram for an audio clip labeled as "angry". The waveform in the top graph displays distinct, spiked patterns indicating strong amplitude variations over time, which suggests an intense and powerful emotional expression. The spectrogram in the bottom graph visualizes the frequency content, with areas of higher intensity and broader frequency ranges, further conveying the agitated and forceful nature of the "angry" audio

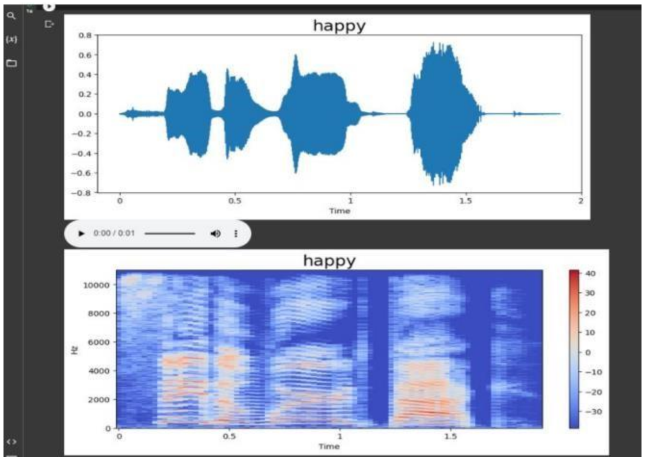

Figure-4.7: Audio Signal for Happy Emotion

Figure 4 The image shows the waveform and spectrogram for an audio clip labeled as "happy". The waveform displays a more varied and undulating pattern compared to the previous "angry" example, with both positive and negative amplitude peaks, suggesting a more upbeat and lively emotional expression. The spectrogram visualizes a broader range of frequencies with areas of higher intensity, conveying a sense of vibrancy and energy associated with the "happy" audio.

The combination of the waveform and spectrogram provides a comprehensive representation of the dynamic and multifaceted qualities inherent in the audio's "happy" emotional expression..

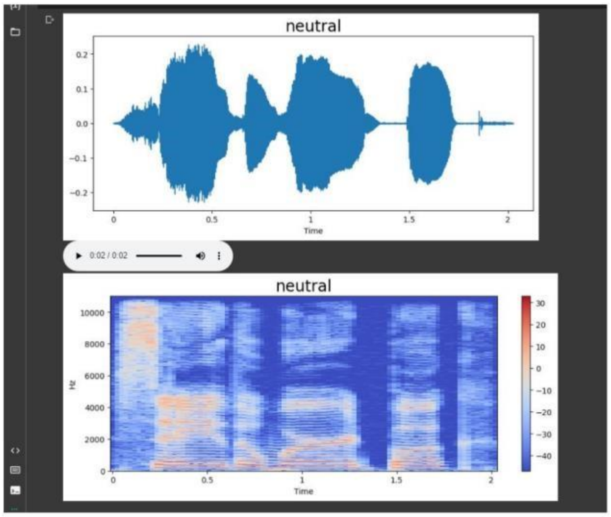

Figure-4.8: Audio Signal for Neutral emotion

Figure 4.8 The image shows the waveform and spectrogram for an audio clip labeled as "neutral". The waveform has a more regular and less pronounced pattern compared to the previous emotional examples, with smaller amplitude fluctuations, suggesting a more balanced, even-keeled emotional expression. The spectrogram also exhibits a more uniform distribution of frequencies without the distinct intensity peaks observed in the "happy" and "angry" examples, further conveying a sense of neutrality and lack of strong emotional content in the audio. The combination of the waveform and spectrogram provides a comprehensive representation of the relatively stable and unemotional qualities inherent in the "neutral" audio clip.

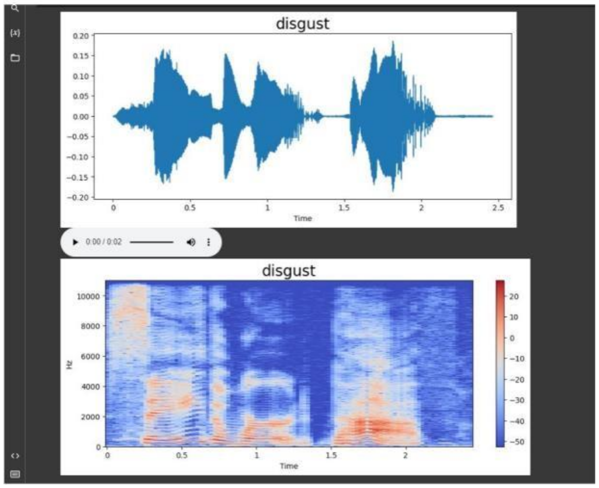

Figure-4.9: Audio Signal for Disgust Emotion

Figure 4 The image shows the waveform and spectrogram for an audio clip labeled as "disgust". The waveform exhibits sharp, irregular peaks and valleys, suggesting a strong, uneven emotional expression. The spectrogram visualizes a broader range of frequencies with distinct areas of high intensity, conveying a sense of complexity and turbulence associated with the "disgust" emotion. The combination of the waveform and spectrogram provides a comprehensive representation of the intense and unsettling qualities inherent in the audio's "disgust".

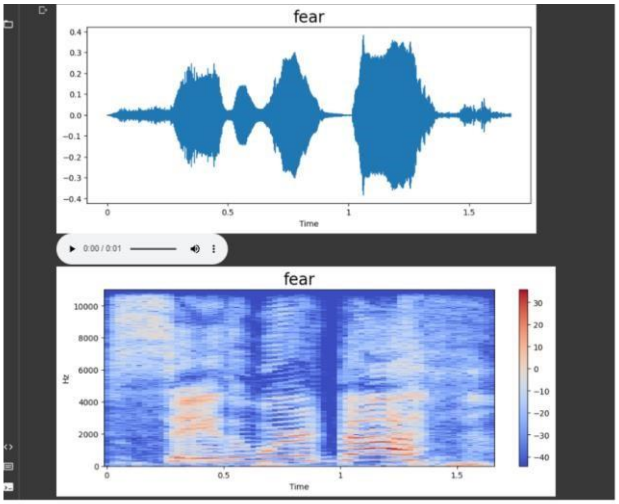

Figure-4.10: Audio Signal for Fear Emotion

Figure 4.10 The image shows the waveform and spectrogram for an audio clip labeled as "fear". The waveform exhibits sharp, irregular peaks and valleys, with a sense of unease and tension in the amplitude variations. The spectrogram visualizes a wider range of frequencies with areas of high intensity, conveying a sense of complexity and turbulence associated with the "fear" emotion. The combination of the waveform and spectrogram provides a comprehensive representation of the intense, unsettling, and restless qualities inherent in the audio's "fear" emotional expression.

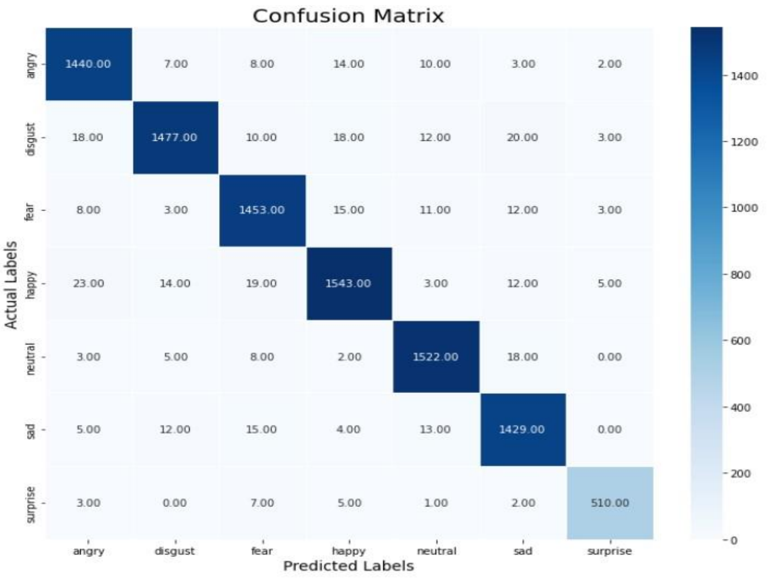

Figure-4.11:Confusion Matrix for CNN

Figure-4.11 The table illustrates the effectiveness of a classification model. The rows denote the actual emotions, while the columns indicate the emotions predicted by the model. The figures within the cells represent the frequency of correct and incorrect predictions. For instance, the top left cell indicates that the model accurately identified 1,440 instances of happiness. Conversely, the bottom right cell reveals that the model mistakenly classified 510 instances of surprise as neutral..

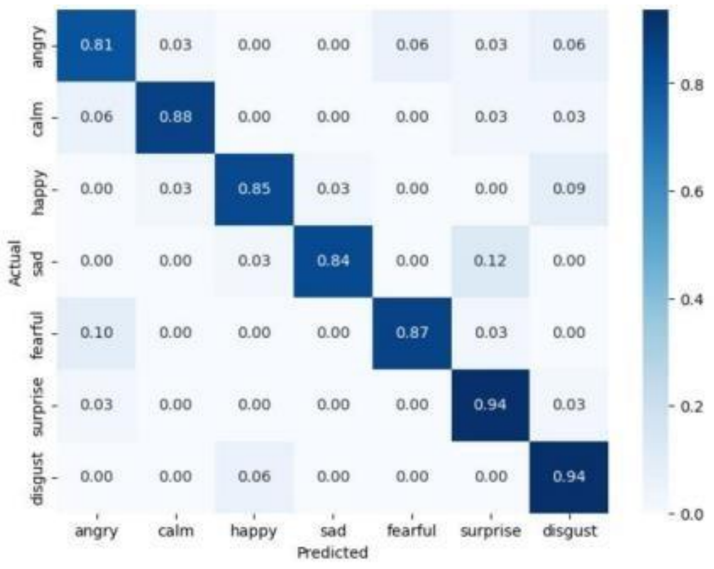

Figure-4.13: Confusion Matrix for LSTM

Figure 4.13 depicts the confusion matrix, which is used to evaluate the performance of an algorithm. In this case, the algorithm is designed to recognize emotions from facial expressions. The rows represent the actual emotions, and the columns represent the emotions that the algorithm predicted. The numbers in the boxes represent the percentage of emotions that were correctly and incorrectly classified. The top left box shows that 81% of the time the algorithm correctly classified angry faces. The bottom right box shows that 94% of the time the algorithm correctly classified calm faces.

CHAPTER-5
CONCLUSION

5.1 CONCLUSION

The Speech Emotion Recognition (SER) project represents a significant endeavor aimed at comprehending and interpreting human emotions conveyed through speech. By utilizing machine learning algorithms and signal processing techniques, SER autonomously extracts and analyzes features from speech data to discern emotional states accurately. Using datasets like the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), SER has uncovered patterns indicative of various emotional states, enriching our understanding of human emotions. The architecture of the SER system, integrating CNNs and layers, effectively extracts features from speech audio data. By improving human-computer interaction, SER presents opportunities for applications across disciplines such as emotional computing, natural language processing, and human-centered technology. Several avenues for future exploration emerge within the realm of Speech Emotion Recognition. The combination of several modalities, such as facial expressions and text data, shows potential for improving emotion recognition accuracy. Developing techniques for real-time emotion recognition would facilitate applications in dynamic environments. Exploring methods for incremental learning could amplify the adaptability and efficacy of the SER system. Conducting cross-cultural studies and tailoring SER models to diverse cultural contexts could lead to more culturally attuned emotion recognition systems. Further inquiry into the ethical implications of SER is imperative for responsible AI development. Pursuing these pathways may pave way for more precise, efficient, and culturally sensitive emotion identification systems.

In this Project we have used the models like CNN, SVM, LSTM For these models we used RAVDESS dataset and Emo-DB dataset. We applied these models for different emotions and got different accuracies. Based on all the accuracies CNN model using RAVDESS dataset got high accuracy that is 96.

CHAPTER-6
REFERENCES

REFERENCES

[1] J. Santoso, K. Ishizuka, and T. Hashimoto, “Large Language Model-Based Emotional Speech Annotation Using Context and Acoustic Feature for Speech Emotion Recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 11026–11030, 2024, doi: 10.1109/ICASSP48485.2024.10448316.

[2] M. Sidorov, S. Ultes, and A. Schmitt, “Emotions are a personal thing: Towards speaker-adaptive emotion recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 4803–4807, 2014, doi: 10.1109/ICASSP.2014.6854514.

[3] J. Zhang, L. Jiang, Y. Zong, W. Zheng, and L. Zhao, “Cross-corpus speech emotion recognition using joint distribution adaptive regression,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, pp. 3790–3794, 2021, doi: 10.1109/ICASSP39728.2021.9414372.

[4] L. Zhang, Y. Wang, J. Du, and X. Wang, “CNN-BiGRU Speech Emotion Recognition Based on Attention Mechanism,” *Proc. - 2023 2nd Int. Conf. Artif. Intell. Intell. Inf. Process. AIIIP 2023*, pp. 85–89, 2023, doi: 10.1109/AIIIP61647.2023.00022.

[5] M. H. Abdul-Hadi and J. Waleed, “Human Speech and Facial Emotion Recognition Technique Using SVM,” *Proc. 2020 Int. Conf. Comput. Sci. Softw. Eng. CSASE 2020*, pp. 191–196, 2020, doi: 10.1109/CSASE48920.2020.9142065.

[6] M. Bhaykar, J. Yadav, and K. S. Rao, “Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM,” *2013 Natl. Conf. Commun. NCC 2013*, pp. 1–5, 2013, doi: 10.1109/NCC.2013.6487998.

[7] M. Sakurai and T. Kosaka, “Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results,” *2021 IEEE 10th Glob. Conf. Consum. Electron. GCCE 2021*, pp. 824–827, 2021, doi: 10.1109/GCCE53005.2021.9621810.

[8] X. Ke, B. Cao, J. Bai, Q. Yu, and D. Yang, “Speech emotion recognition based on PCA and CHMM,” *Proc. 2019 IEEE 8th Jt. Int. Inf. Technol. Artif. Intell. Conf. ITAIC 2019*, no. Itaic, pp. 667–671, 2019, doi: 10.1109/ITAIC.2019.8785867.

[9] B. Chen, Q. Yin, and P. Guo, “A study of deep belief network based Chinese speech emotion recognition,” *Proc. - 2014 10th Int. Conf. Comput. Intell. Secur. CIS 2014*, pp. 180–184, 2015, doi: 10.1109/CIS.2014.148.

[10] A. Firoz Shah and P. Babu Anto, “Hybrid spectral features for speech emotion recognition,” *Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017*, vol. 2018-January, pp. 1–4, 2017, doi: 10.1109/ICIIECS.2017.8275943.

[11] Y. X. Xi, Y. Song, L. R. Dai, I. McLoughlin, and L. Liu, “Frontend Attributes Disentanglement for Speech Emotion Recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, pp. 7712–7716, 2022, doi: 10.1109/ICASSP43922.2022.9746691.

[12] Y. Li, P. Bell, and C. Lai, “Fusing Asr Outputs in Joint Training for Speech Emotion

Recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, pp. 7362–7366, 2022, doi: 10.1109/ICASSP43922.2022.9746289.

[13] I. R. Ulgen, Z. Du, C. Busso, and B. Sisman, "Revealing Emotional Clusters in Speaker Embeddings: A Contrastive Learning Strategy for Speech Emotion Recognition," *ICASSP 2024 - 2024 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 12081–12085, 2024, doi: 10.1109/icassp48485.2024.10447060.

[14] P. Yue, L. Qu, S. Zheng, and T. Li, "Multi-task Learning for Speech Emotion and Emotion Intensity Recognition," *Proc. 2022 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2022*, no. November, pp. 1232–1237, 2022, doi: 10.23919/APSIPAASC55919.2022.9979844.

[15] S. Shen, Y. Gao, F. Liu, H. Wang, and A. Zhou, "Emotion Neural Transducer for Fine-Grained Speech Emotion Recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 10111–10115, 2024, doi: 10.1109/ICASSP48485.2024.10446974.

[16] R. Sato, R. Sasaki, N. Suga, and T. Furukawa, "Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition," *Proc. 2020 23rd Conf. Orient. COCOSDA Int. Comm. Co-ord. Stand. Speech Databases Assess. Tech. O-COCOSDA 2020*, pp. 33–37, 2020, doi: 10.1109/O-COCOSDA50338.2020.9295041.

[17] D. S. Moschona, "An Affective Service based on Multi-Modal Emotion Recognition, using EEG enabled Emotion Tracking and Speech Emotion Recognition," *2020 IEEE Int. Conf. Consum. Electron. - Asia, ICCE-Asia 2020*, 2020, doi: 10.1109/ICCE-Asia49877.2020.9277291.

[18] R. Chakraborty, M. Pandharipande, and S. K. Kopparapu, "Spontaneous speech emotion recognition using prior knowledge," *Proc. - Int. Conf. Pattern Recognit.*, vol. 0, pp. 2866–2871, 2016, doi: 10.1109/ICPR.2016.7900071.

[19] N. Jia and C. Zheng, "Emotion Recognition of Depressive Patients Based on General Speech Information," *2021 IEEE 6th Int. Conf. Intell. Comput. Signal Process. ICSP 2021*, no. Icsp, pp. 618–621, 2021, doi: 10.1109/ICSP51882.2021.9408759.

[20] Wang, H., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. IEEE Transactions on Multimedia, 10(5), 936-946.