A project report on

EMOTION DETECTION THROUGH VOICE

Submitted in partial fulfillment for the award of the degree of

MASTER OF SCIENCE

IN

DATA SCIENCE

By

VAISHNAVI R (23MSD7028)

VIT-AP UNIVERSITY

SCHOOL OF ADVANCED SCIENCES

VIT-AP UNIVERSITY

AMARAVATI- 522237

MAY, 2025

EMOTION DETECTION THROUGH VOICE

Submitted in partial fulfillment for the award of the degree of

MASTER OF SCIENCE

IN

DATA SCIENCE

By

VAISHNAVI R (23MSD7028)

SCHOOL OF ADVANCED SCIENCES

VIT-AP UNIVERSITY

AMARAVATI- 522237

MAY, 2025

DECLARATION

I hereby declare that the thesis entitled “EMOTION DETECTION THROUGH VOICE” submitted by me, for the award of the degree of MSc Data Science VIT is a record of bonafide work carried out by me under the supervision of Jay Kumar.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Amaravati

Date: 15 May 2025 Signature of the Candidate

NULLCLASS

CERTIFICATE

OF INTERNSHIP

THIS CERTIFICATE IS PRESENTED TO

Vaishnavi R

has successfully completed his/her Data Science internship from Dec 12 2024 till Apr 12 2025.

He/She has participated successfully in all tasks given to him/her and accomplished all the skills required for the tasks,throughout which he/she was able to showcase his/her great work ethics and team player skills.

Issued On: Apr 21 2025
GSTIN: 33AAICN0803F1ZZ
CIN: U80301TZ2022PTC038231
DOC ID: 6805b5c4da8c290a06240026

Vetriselvan G.
CEO

Verfiy at:

CERTIFICATE

This is to certify that the Internship titled "**EMOTION DETECTION THROUGH VOICE**" that is being submitted by **VAISHNAVI R (23MSD7028)** is in partial fulfillment of the requirements for the award of Master of Science in Data Science, is a record of bonafide work done under my guidance. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for award of any degree or diploma and the same is certified.

Mr. Jay Kumar

External Guide

(with company seal)

The thesis is satisfactory / unsatisfactory

Internal Examiner **External Examiner**

Approved by

PROGRAM CHAIR
MSc Data Science

DEAN
School of Advanced Sciences

ABSTRACT

Speech Emotion Recognition (SER) systems have significant real-world applications in areas such as virtual assistants, mental health monitoring, call center analytics, and human-computer interaction, where understanding emotional context can enhance responsiveness and user experience. This project Emotion Detection through Voice aims to develop a deep learning-based SER system capable of classifying emotions from voice recordings.

The system is trained on the CREMA-D dataset, which includes a wide range of labeled speech samples representing six core emotions: Angry, Disgust, Fear, Happy, Neutral, and Sad. Audio preprocessing steps such as waveplot and spectrogram visualization were conducted, followed by feature extraction using Mel Frequency Cepstral Coefficients (MFCCs). To increase robustness, audio data was augmented through noise addition, pitch shifting, time stretching, and shifting.

A hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) layers was implemented to capture both spatial and temporal features in the audio. The trained model achieved a validation accuracy of approximately 84%. A user-friendly Streamlit-based web interface was also developed, allowing users to upload or record audio in real time for emotion prediction. This project demonstrates the potential of integrating deep learning with real-time applications for emotion-aware systems.

AKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to JAY KUMAR, DATA SCIENCE PROJECT HEAD AND MANAGER NULL CLASS for his constant guidance, continual encouragement, understanding; more than all, he taught me patience in my endeavor. My association with him is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of DATA SCIENCE.

I would like to express my gratitude to Dr. G. Viswanathan, Dr. Sekar Viswanathan, Dr. S. V. Kota Reddy, and Dr. S. Srinivas, School of Advanced Sciences (SAS), for providing with an environment to work in and for his inspiration during the tenure of the course.

In jubilant mood I express ingeniously my whole-hearted thanks to Dr. Faizan Danish, Program Chair of the Data Science Department, and Dr. Vemula Ramakrishna Reddy, Head of the Mathematics Department,

all teaching staff and members working as limbs of our university for their notself-centered enthusiasm coupled with timely encouragements showered on me with zeal, which prompted the acquirement of the requisite knowledge to finalize my course study successfully. I would like to thank my parents for their support.

It is indeed a pleasure to thank my friends who persuaded and encouraged me to take up and complete this task. At last but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly toward the successful completion of this project

Place: Amaravati

Date:15 May 2025

Name of the student

Vaishnavi R

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AI Artificial Intelligence
API Application Programming Interface
CREMA-D Crowd-sourced Emotional Multimodal Actors Dataset
CNN Convolutional Neural Network
GCN Graph Convolutional Network
GUI Graphical User Interface
HCI Human-Computer Interaction
LSTM Long ShortTerm Memory
MFCC Mel-Frequency Cepstral Coefficients
ML Machine Learning
NLP Natural Language Processing
RNN Recurrent Neural Network
SER Speech Emotion Recognition
UI User Interface

CHAPTER 4

METHODOLOGY

CHAPTER 5

EXPERIMENTS AND RESULTS

CHAPTER 1

INTRODUCTION

1.1 EMOTION DETECTION THROUGH VOICE

Emotion detection through voice is a rapidly evolving field that involves the automatic identification and interpretation of human emotions from spoken language using computational techniques. Unlike written or textual communication, spoken language conveys a rich set of emotional signals beyond the literal meaning of words. These signals known as paralinguistic features include pitch, tone, speech rate, rhythm, volume, and intensity, which are often unconsciously modulated depending on a speaker's emotional state.

The ability to analyze these acoustic and prosodic features allows machines to infer emotional conditions such as happiness, sadness, anger, fear, disgust, surprise, and more. This process is known as Speech Emotion Recognition (SER) and forms a critical subdomain of affective computing, which aims to create systems that can recognize, interpret, and respond to human emotions in a natural and intelligent way.

With advances in machine learning and deep learning, particularly through techniques such as feature extraction (e.g., MFCCs) and neural network-based classification, SER systems have achieved significant improvements in accuracy and real-time performance. These systems are now being applied in a wide range of applications, including human-computer interaction, customer service, healthcare diagnostics, education, surveillance, and virtual assistants.

By enabling machines to perceive and appropriately react to human emotional states, emotion detection through voice plays a pivotal role in enhancing user experience, personalizing responses, and making technology more empathetic and human-centric.

1.2 MOTIVATION FOR THE EMOTION DETECTION THROUGH VOICE PROJECT

In recent years, there has been a rapidly growing demand for emotionally intelligent systems across a wide range of domains, including healthcare, customer service, education, and entertainment. In healthcare, emotion recognition systems can assist in the early detection and continuous monitoring of mental health conditions such as depression, anxiety, and emotional distress, providing valuable support for clinicians and caregivers. In customer service and call center environments, understanding the emotional state of a customer can help agents respond more empathetically, leading to improved customer satisfaction and resolution efficiency. In the field of education, emotion-aware e-learning platforms can adapt content delivery based on students' emotional states, promoting better engagement and learning outcomes. Entertainment platforms and virtual assistants can also benefit from emotion-sensitive responses, enhancing user interaction and personalization.

Recognizing these possibilities, the motivation behind this project was to develop a Speech Emotion Recognition (SER) system that goes beyond academic experimentation and serves real-world use cases. Human speech carries a wealth of emotional cues, and accurately identifying them requires sophisticated techniques capable of handling variations in tone, pitch, pace, and expression. Therefore, the goal was not only to build a highly accurate classification model but also to make it resilient, scalable, and easy to use in dynamic environments. Leveraging recent advances in deep learning, particularly in Convolutional and Recurrent Neural Networks, this project aims to classify emotions from speech with high precision while addressing practical challenges such as background noise and emotional overlap.

Moreover, to ensure the model's applicability in real-world scenarios, a user-friendly Graphical User Interface (GUI) was developed. This interface allows users to either upload pre-recorded audio or record live speech, making the system accessible and interactive for non-technical users. The overall motivation is rooted in the belief that emotionally aware technologies will play a pivotal role in shaping future human-computer interactions, offering systems that are not just intelligent—but also empathetic and context-aware.

1.3 CHALLENGES PRESENT IN EMOTION DETECTION THROUGH VOICE

Developing a reliable and accurate speech emotion recognition (SER) system involves addressing a range of technical and real-world challenges. One of the primary difficulties lies in the quality and diversity of audio data. In real-world settings, people speak in varied tones, speeds, volumes, and emotional intensities. Additionally, accents, gender differences, age, and cultural variations all influence how emotions are expressed vocally. These variations make it challenging for models to generalize effectively across different speakers and environments.

Another significant challenge is emotion overlap. Emotions such as fear and sadness or anger and disgust often exhibit similar acoustic patterns, such as comparable pitch or energy levels, which can confuse even well-trained models. These emotional categories are not always distinct in practice, and sometimes a single utterance may contain cues from multiple emotional states, further complicating classification.

Background noise is another critical obstacle, especially in real-time or real-world deployment scenarios. Audio recordings may contain ambient sounds like traffic, background chatter, electronic hums, or environmental disturbances, which can degrade the quality of extracted features. Models trained on clean data may struggle to perform under such noisy conditions unless properly augmented or denoised.

Moreover, there is the constant risk of overfitting, particularly when working with deep learning architectures like CNNs or LSTMs that contain a large number of trainable parameters. If the model is exposed to a limited or unbalanced dataset, it may memorize patterns specific to the training data rather than learning generalizable representations. This results in poor performance on unseen test data. Addressing overfitting requires techniques such as data augmentation, dropout layers, validation strategies, and early stopping during training.

In summary, speech emotion recognition demands careful consideration of data variability, emotional ambiguity, environmental noise, and model generalization to ensure that the system performs robustly in diverse and practical applications.

1.4 SCOPE OF PROJECT

The scope of this project is centered on building a complete, end-to-end Speech Emotion Recognition (SER) system that can accurately classify emotional states from human speech using deep learning techniques. It begins with the use of the CREMA-D dataset, which offers a diverse range of audio samples labeled with six distinct emotions such as angry, Disgust, Fear, Happy, Neutral, and Sad. The project includes thorough preprocessing of the audio data, where emotion labels are extracted from file names and a structured dataset is created. It then involves visual analysis of speech signals through waveplots and spectrograms to understand time-domain and frequency-domain characteristics of different emotions. Core to the project is the extraction of relevant acoustic features using Mel Frequency Cepstral Coefficients (MFCCs), which translate raw audio into machine-understandable numerical representations. To expand the dataset's diversity and enhance the model's generalization, various data augmentation techniques such as noise addition, pitch shifting, time stretching, and shifting are applied. A robust deep learning architecture combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks is designed to capture both local audio features and temporal patterns over time. The model is trained and validated using standard evaluation metrics—accuracy, precision, recall, and F1-score—to assess its predictive performance across emotion classes. Additionally, the scope extends to practical deployment through the development of an interactive web-based Graphical User Interface (GUI) using Streamlit, enabling real-time emotion prediction from both uploaded and live-recorded audio. Overall, the project encompasses the full cycle of data handling, model development, performance evaluation, and application deployment within the field of emotion recognition from speech.

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE REVIEW

a) Temporal aggregation of audio-visual modalities for emotion recognition

Methodology:

The paper proposes a multimodal fusion technique for emotion recognition based on combining audio-visual modalities.The fusion is done using temporal aggregation with different temporal offsets for each modality (audio and video), allowing asynchronous inputs.

- The process involves:
- Dividing the video into temporal segments.
- Randomly selecting a video frame and corresponding audio segment for each segment.
- Using a core audio-visual network composed of convolutional blocks to extract features from both modalities.
- Aggregating the emotion probabilities for all temporal segments to predict the final emotion.

Dataset Used:

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):

The dataset consists of 7442 clips from 91 actors (48 male, 43 female) who were instructed to express six different emotions (neutral, happy, anger, disgust, fear, sad) using 12 sentences. Each clip was labeled with emotions and intensity levels.The dataset also includes audio-visual data for emotion recognition tasks.

Evaluation Metrics:

- Accuracy: The mean accuracy across all the test folds.
- Loss: Cross-entropy loss function used for training.
- Confusion Matrix: Used to evaluate the performance of the model by showing the true vs predicted emotions.
- Training and Validation Loss/Accuracy: Performance on both sets.

- In their experiments:
- They reported the human accuracy for emotion recognition in CREMA-D as 63.6%.
- Their method achieved 68.4% accuracy, outperforming several existing methods like CNN-based approaches (55.8%) and Recurrent Multi-Attention (RMA) (65%).

Findings:

The proposed asynchronous fusion of audio and video data outperforms previous models in terms of accuracy.The temporal aggregation mechanism effectively combines information from both modalities, leading to improved performance.

They found that adding more segments for temporal aggregation increased accuracy but also increased training and inference time. However, accuracy improvements plateaued after 8 segments.The core audio-visual network architecture, featuring convolutional blocks, effectively extracted discriminative features from both modalities, resulting in high emotion recognition accuracy.

Relevance to the Project:

The paper's approach to combining audio and video modalities asynchronously through temporal aggregation is highly relevant to the Emotion Detection through voice project, which already uses audio-based features.

Although the project focuses on speech emotion recognition using audio features (MFCCs, spectrograms) and a CNN+LSTM model, this paper offers insights on how combining audio and visual data could potentially enhance the recognition accuracy.

Could explore incorporating visual features (e.g., face detection, expression analysis) from video data to complement the audio-based emotion recognition in the project.

Their core CNN architecture for extracting audio-visual features could inspire in designing or enhancing the deep learning network for the SER project**.**

b) Multimodal Emotion Recognition Using Audio-Visual Features and Graph Convolutional Networks

Methodology:

The paper proposes a multimodal emotion recognition framework that combines both audio and visual features using Graph Convolutional Networks (GCN). The framework utilizes a feature fusion approach where the audio and visual features are extracted and

then processed separately using convolutional layers before being fused. The GCN is employed to capture the temporal dependencies between the modalities and perform the emotion recognition task.

Key steps in the methodology:

1. Audio Feature Extraction: Spectrograms of the audio signals are extracted using Short-Time Fourier Transform (STFT) and processed using Convolutional Neural Networks (CNNs).
2. Visual Feature Extraction: The facial expressions are extracted using a pre-trained CNN model like VGG-16, which is fine-tuned for emotion recognition.
3. Graph Convolutional Network (GCN): A graph structure is formed to model the interactions between audio-visual features. GCN is used to capture temporal dependencies in the multimodal data.
4. Fusion: The audio and visual features are concatenated and passed through a GCN model to enhance the temporal aggregation of the features.
5. Emotion Classification: The final emotion category is predicted based on the fused audio-visual features.

Dataset Used:

CREMA-D - A multimodal dataset that includes 7442 clips of 91 actors conveying different emotions through speech and facial expressions. The dataset contains six emotions: neutral, happy, anger, disgust, fear, and sad.

Evaluation Metrics:

- Accuracy: The performance is evaluated based on the overall accuracy of emotion classification.
- Confusion Matrix: Used to assess the classification performance for each emotion category.
- Precision, Recall, and F1-Score: These metrics are used to measure the performance for individual emotion classes.
- Mean Accuracy: Evaluated using 10-fold cross-validation.

Findings:

- The proposed model significantly outperforms traditional methods that rely solely on audio or visual features.
- The GCN-based fusion of audio-visual features leads to a more accurate emotion recognition system.

- The method achieved an accuracy of 79.1%, outperforming previous models that utilized only either audio or visual modalities.
- The performance of the proposed approach is robust across various emotion categories, showing higher classification accuracy for emotions like happiness and anger.

Relevance to the Project:

This paper is highly relevant to the emotion detection though voice project, as it proposes a multimodal approach combining both audio and visual features for emotion recognition. Since the project is working with the CREMA-D dataset, it can adapt some of the techniques in this paper, especially the feature extraction from both speech and facial expressions. The integration of Graph Convolutional Networks (GCN) for capturing temporal dependencies could improve the model's performance, particularly in your CNN+LSTM architecture, which might also benefit from GCN's ability to model interactions between modalities over time. This approach could be used to further enhance the emotion detection capabilities of the model by combining both the visual and speech information in a more structured and temporal way.

2.2 SURVEY ON EMOTION DETECTION THROUGH VOICE IN AI

Speech Emotion Recognition (SER) is a critical domain within AI that focuses on detecting and interpreting emotions from speech signals. This technology has been increasingly applied across various industries. Below, we focus on two main applications of SER in AI:

2.2.1. HUMAN-COMPUTER INTERACTION (HCI)

Human-Computer Interaction is one of the most significant applications of SER. In this domain, speech emotion recognition helps in improving the quality of user interactions with machines. By understanding human emotions through speech, machines can respond in a way that feels more natural and empathetic, which is essential for virtual assistants (like Siri or Alexa), customer service bots, and gaming systems. SER allows these systems to recognize frustration, joy, or anger in a user’s tone, adjusting their responses accordingly.

Applications in HCI:

- Virtual Assistants: Enhances user interaction by recognizing the emotional state of the user, providing empathetic responses.
- Customer Support Systems: Helps customer service bots detect customer emotions such as frustration or satisfaction, enabling better responses and escalation protocols.
- Entertainment & Gaming: Provides emotional feedback in gaming environments where players' emotional states are tracked to adjust gameplay experiences.

2.2.2 HEALTHCARE AND THERAPY

SER has emerged as a promising tool in the healthcare sector, particularly for mental health monitoring. Detecting emotions such as stress, anxiety, or depression through speech is valuable for mental health professionals, especially in telemedicine. By analyzing voice patterns, therapists can better understand a patient's emotional well-being over time, even in remote consultations.

Applications in Healthcare:

- Mental Health Monitoring: Detects emotional states such as sadness, anxiety, or depression in patients, providing early signs for further intervention.
- Assistive Technologies: Helps patients with speech disabilities or those who are unable to communicate effectively by analyzing their emotional states, thus supporting therapy.
- Telemedicine and Remote Monitoring: Enables remote healthcare professionals to assess patients' emotional well-being without physical visits.

2.3 EXISTING SYSTEMS AND TECHNOLOGIES IN EMOTION DETECTION THROUGH VOICE

Several software systems and frameworks support the development of Speech Emotion Recognition systems by providing feature extraction tools, pre-built models, and interfaces for training and evaluation. These include:

OpenSMILE (Speech and Music Interpretation by Large Space Extraction):

OpenSMILE is one of the most widely used toolkits for speech feature extraction. It can extract a wide variety of audio features, including MFCCs, pitch, energy, and other prosodic features, making it ideal for emotion recognition.

TensorFlow and PyTorch:
These deep learning frameworks are commonly used to build custom SER models. TensorFlow and PyTorch support the implementation of complex neural network architectures like CNN, LSTM, and hybrid models (CNN + LSTM), which are crucial for accurate emotion classification.

Kaldi:
Kaldi is an open-source speech recognition toolkit that provides tools for feature extraction, acoustic modeling, and training machine learning models. It's widely used in speech recognition and emotion recognition tasks.

EmoReact:
EmoReact is a multimodal emotion recognition system that integrates speech, visual, and textual data for more robust emotion classification. While primarily used for multimodal systems, its speech recognition capabilities are useful for building emotion detection models based on voice.

DeepSBD (Deep Speech-Based Emotion Detection):
DeepSBD uses deep learning techniques to analyze speech data and classify emotions. It often employs CNNs or RNNs for classification, based on preprocessed audio features like MFCCs or spectrograms.

Microsoft Azure Speech Service and Google Cloud Speech-to-Text:
These cloud-based APIs provide tools for speech recognition and can be extended for emotion detection by adding emotion classification layers on top of recognized speech. They offer the advantage of easy integration into real-time applications and cloud-based deployment.

SpeechBrain:
SpeechBrain is an open-source PyTorch-based toolkit designed for speech processing tasks. It includes pre-trained models for speech emotion recognition, which can be fine-tuned on custom datasets.

Some platforms integrate the entire SER pipeline, including data preprocessing, model training, and inference, making it easier to develop and deploy SER systems:

IBM Watson Tone Analyzer:
IBM Watson provides a Tone Analyzer service that can detect emotions like joy, sadness, and anger from speech or text. It uses a combination of linguistic and acoustic features for emotion recognition.

iFLYTEK Emotion Recognition API:
iFLYTEK offers an emotion recognition API that analyzes speech for emotional states such as happiness, sadness, and anger. It is widely used in customer service, health, and entertainment applications.

Sensory TrulyHandsfree:
This platform focuses on speech recognition and emotion detection through a combination of acoustic modeling and natural language processing. It can be used in voice assistants, security systems, and interactive devices.

CHAPTER 3

BACKGROUND AND RELATED WORK

3.1 INTRODUCTION

This chapter provides a theoretical foundation for the field of Speech Emotion Recognition (SER), offering insight into its psychological, computational, and technological underpinnings. The discussion traces the evolution of SER technologies, identifies commonly used emotion models, explores the nature of speech as an emotional signal, and reviews key research contributions in this domain. The objective is to contextualize the current work within the broader research landscape and highlight the scientific principles that inform modern SER systems. This foundation sets the stage for the methodological framework presented in the next chapter.

3.2 FOUNDATIONS OF EMOTION DETECTION THROUGH VOICE

Speech Emotion Recognition is a vital subfield of affective computing, which aims to enable machines to recognize and respond to human emotions. Unlike facial expressions or physiological signals, speech is a particularly rich and accessible medium for emotion expression. SER systems analyze vocal cues to determine the emotional state of a speaker, enabling more empathetic and adaptive interactions in human-computer interfaces.

The evolution of SER began with rule-based and statistical models in the early 2000s, where features like pitch and energy were manually crafted and input into classifiers such as Support Vector Machines (SVM) or Hidden Markov Models (HMM). Over time, the field transitioned to machine learning approaches and, more recently, to deep learning models, which can automatically learn discriminative features from raw or minimally processed audio inputs.

3.3 EMOTION MODELS IN SER

Understanding how emotions are represented and categorized is essential to building SER systems. Two primary types of models are commonly used:

Categorical Emotion Models

Categorical models classify emotions into a finite set of discrete categories such as:

- Anger
- Happiness
- Sadness
- Fear
- Disgust
- Surprise

This approach aligns well with how many datasets (like CREMA-D, RAVDESS, TESS, and SAVEE) are annotated. It is simple to implement and interpret, making it ideal for practical applications.

Dimensional Emotion Models

These models represent emotions as points in a continuous multi-dimensional space. The most popular is Russell's Circumplex Model of Affect, which uses:

- Valence (positive vs. negative emotions)
- Arousal (intensity of emotion)

This model is more expressive and is better suited for nuanced emotional states (e.g., calm happiness vs. excited happiness). However, collecting annotated data for these models is more complex and subjective.

3.4 ACOUSTIC FEATURES IN EMOTION RECOGNITION

Speech conveys emotion not through the words themselves, but through how they are spoken—tone, rhythm, energy, etc. The acoustic features used in SER are broadly classified as follows:

Prosodic Features

- Pitch (F0): Higher or lower pitch can indicate emotions like happiness or sadness.
- Energy (Loudness): Angry speech is often louder; sad speech tends to be soft.
- Duration and Speaking Rate: Fast speech may indicate excitement or nervousness, while slow speech may signal sadness or fatigue.

Spectral Features

These include:

- Mel-Frequency Cepstral Coefficients (MFCCs) are among the most widely used features in speech and audio processing, particularly in Speech Emotion Recognition, due to their ability to effectively capture the timbral and spectral characteristics of the human voice. MFCCs are derived by first transforming an audio signal into a power spectrum using the Short-Time Fourier Transform (STFT), which breaks the signal into small overlapping frames to analyze its frequency content over time. The power spectrum is then passed through a set of filters spaced according to the Mel scale, a perceptual scale that approximates how the human ear perceives pitch, placing more emphasis on lower frequencies where the ear is more sensitive. After filtering, the logarithm of the energy in each Mel band is taken to simulate the logarithmic perception of loudness, and finally, a Discrete Cosine Transform (DCT) is applied to these log energies to reduce dimensionality and decorrelate the features. The result is a set of coefficients, the MFCCs, that compactly represent the shape of the vocal tract during speech production, making them highly effective for distinguishing between different emotional states, as emotions alter vocal tract configurations and excitation patterns.
- Chroma Features: Capture energy in each pitch class.
- Spectral Centroid, Bandwidth, and Rolloff: Describe where energy is concentrated in the frequency spectrum.

Voice Quality Features

- Jitter and Shimmer: Variations in frequency and amplitude that reflect emotional tension or instability in the voice.
- Harmonics-to-Noise Ratio (HNR): Measures the ratio of harmonic sound to noise, helpful in detecting vocal strain or breathiness.

These features serve as the input for traditional machine learning models and provide interpretability. Deep learning models may use them or learn directly from raw audio signals.

3.5 RELATED WORK IN EMOTION DETECTION THROUGH VOICE

A substantial body of research has explored various techniques for recognizing emotions from speech. The development can be categorized into three broad eras:

Traditional Machine Learning Approaches

Earlier studies focused on hand-crafted features extracted using signal processing techniques. These features were then input into models such as:

- Support Vector Machines (SVM)
- Hidden Markov Models (HMM)
- Gaussian Mixture Models (GMM)

These models performed reasonably well but often lacked robustness across different speakers, accents, and recording environments.

Deep Learning-Based Approaches

Recent advances leverage deep learning for automated feature extraction and classification:

- Convolutional Neural Networks (CNN): Convolutional Neural Networks (CNNs) are particularly effective in extracting spatial features from two-dimensional data representations like spectrograms or MFCCs, which are commonly used in Speech Emotion Recognition. In SER tasks, speech signals are first transformed into spectrograms or MFCCs—matrix-like representations of the frequency content over time. CNNs apply convolutional filters across these matrices to learn hierarchical patterns that reflect emotional cues such as variations in pitch, tone, or intensity. These spatial patterns often correlate with specific emotional states, making CNNs a powerful tool for learning discriminative features automatically without relying on manual feature engineering. CNNs also offer the benefit of parameter sharing and spatial locality, leading to efficient training even with high-dimensional input data.
- Recurrent Neural Networks (RNN), especially LSTMRecurrent Neural Networks (RNNs), and particularly Long Short-Term Memory (LSTM) networks, are well-suited for modeling temporal dependencies in sequential data like speech. In the context of SER, LSTMs are valuable because emotions are not always expressed instantaneously—they evolve over the duration of an

utterance. LSTMs possess internal memory cells and gating mechanisms that allow them to retain important information across time steps while forgetting irrelevant details. This makes them capable of understanding how emotional states transition and fluctuate throughout a sentence or phrase. Unlike CNNs, which focus on spatial structure, LSTMs emphasize the sequential nature of speech, enabling better emotion tracking over time.

- CNN+LSTM (CLSTM): The hybrid CNN+LSTM architecture (CLSTM) leverages the strengths of both CNNs and LSTMs by combining spatial and temporal learning in a unified framework. In this setup, CNN layers are used first to extract high-level spatial features from spectrogram or MFCC inputs. These extracted features, which may represent pitch contours, energy bursts, or other spectral patterns, are then passed to LSTM layers that model the temporal evolution of these features across the audio sequence. This two-stage approach enables the network to understand both what features are present (spatial) and how they change over time (temporal), leading to more accurate and robust emotion classification. The integration of CNNs and LSTMs has been shown in many studies to outperform models using either architecture alone.

These models significantly improve generalization and performance, especially when trained on diverse and augmented datasets.

Emerging Architectures

- Attention Mechanisms: Focus on emotionally salient parts of the speech signal.
- Transformer-Based Models: Originally developed for NLP, they are being adapted for SER tasks due to their ability to model long-term dependencies without recurrence.
- Pretrained Models (e.g., wav2vec, HuBERT): Provide strong audio representations and can be fine-tuned for emotion classification tasks.

Multimodal and Hybrid Systems

Some advanced systems combine speech with other modalities such as facial expressions (video), physiological signals (like EEG), or textual sentiment (from transcripts) for more accurate emotion detection. While multimodal systems offer improved performance, they require more complex setups and data collection.

CHAPTER 4

METHODOLOGY

4.1 INTRODUCTION

This chapter presents the comprehensive methodology employed in the development of the Speech Emotion Recognition (SER) system using the CREMA-D dataset. The objective of this project is to accurately identify human emotions from audio recordings by leveraging deep learning techniques. The chapter elaborates on each phase of the workflow, beginning with a description of the dataset used and the initial preprocessing steps performed to clean and standardize the audio data. To improve model robustness and generalizability, various audio augmentation techniques such as noise injection, pitch shifting, time stretching, and temporal shifting were applied. Following this, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from the audio samples, which serve as the primary input features for the model due to their proven effectiveness in capturing speech characteristics that correlate with emotional states. The extracted MFCC features were then normalized and reshaped to prepare them for input into the deep learning model. The architecture adopted for this system is a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, also known as CLSTM. This approach is designed to take advantage of the CNN's ability to capture spatial patterns in the audio signal and the LSTM's strength in learning temporal dependencies, which are critical in speech emotion recognition. The final sections of this chapter describe the training strategy, including optimization techniques, evaluation metrics, and model validation. Together, these methodological steps form a robust pipeline for developing a high-performing SER system capable of classifying emotions from speech signals with significant accuracy.

4.2 DATASET DESCRIPTION

The project uses the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) obtained from online resources, is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The

sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

4.3 DATA PREPROCESSING

Data preprocessing is a crucial step in ensuring that the raw audio data is properly formatted and prepared for use in training the Emotion Recognition model. The preprocessing pipeline for this project includes several key steps to standardize the dataset and make it suitable for model input.

4.3.1 FORMAT CONVERSION AND SAMPLING RATE STANDARDIZATION:

The raw audio files in the dataset are often recorded in various formats and sample rates. To ensure consistency, all audio files are converted to a uniform format (typically .wav) and resampled to a standard sampling rate of 22050 Hz. This ensures that all audio files have the same temporal resolution, preventing discrepancies that could impact model performance.

4.3.2 TRIMMING SILENCE:

Silence at the beginning and end of audio recordings is often irrelevant for emotion detection and can introduce unnecessary noise into the data. Therefore, the audio files are processed to automatically detect and trim any silence from both ends. This step ensures that the model focuses on the relevant parts of the speech, such as emotional cues, and reduces the amount of irrelevant data.

4.3.3 STANDARDIZING FILE DURATION:

Audio files in the dataset can vary in length, which could introduce challenges when training the model. Shorter clips are typically zero-padded to a fixed duration, ensuring that all input data has the same length. Zero-padding does not affect the core content of the audio but ensures that the model receives consistent input dimensions.

4.3.4 CONVERTING STEREO TO MONO:

Stereo recordings contain two audio channels (left and right), which can complicate processing. Since the model is primarily concerned with the audio content itself rather

than spatial attributes, all stereo files are converted to mono (single channel). This reduces complexity and ensures that the model focuses solely on the audio features.

4.3.5 WAVEPLOT AND SPECTROGRAM VISUALIZATION:

To gain an initial understanding of the audio characteristics and validate the preprocessing steps, waveplots and spectrograms are generated for each audio clip. Waveplots display the raw amplitude over time, while spectrograms visualize the frequency content over time. These visualizations allow for a quick inspection of the data, helping to identify potential issues such as noise or inconsistencies in pitch. They also provide a clearer view of the emotional content, as different emotions often correspond to distinctive patterns in both the amplitude and frequency domain.

The wave plot represents the amplitude (loudness) of the audio signal over time.The x-axis represents time, while the y-axis represents amplitude.The waveform shows clear variations in amplitude, indicating changes in loudness.

The spectrogram represents how the frequency content of the audio changes over time. The x-axis represents time, while the y-axis represents frequency (Hz).The color intensity indicates the energy at different frequencies: Red/orange areas indicate high energy. Blue areas indicate low energy.

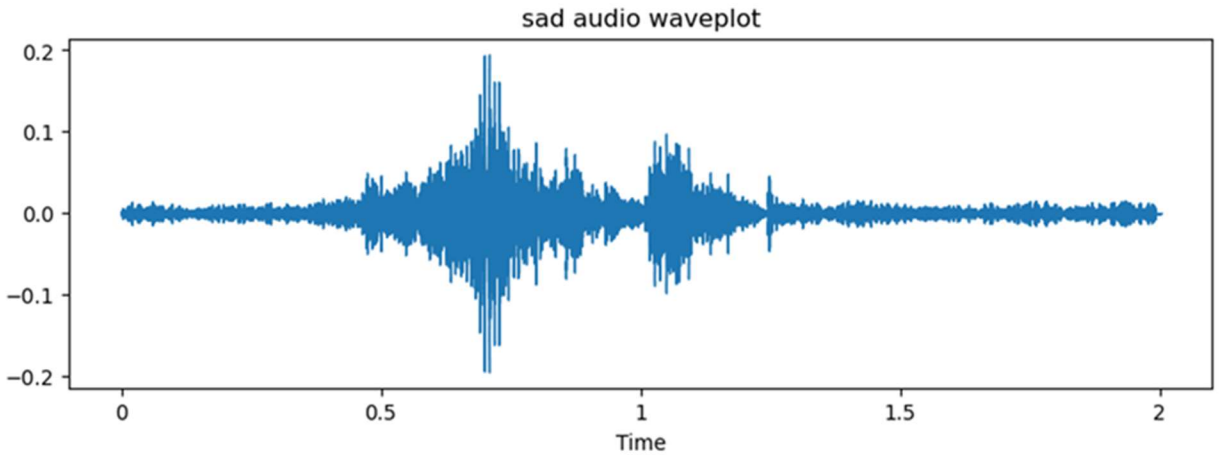

Fig. 1

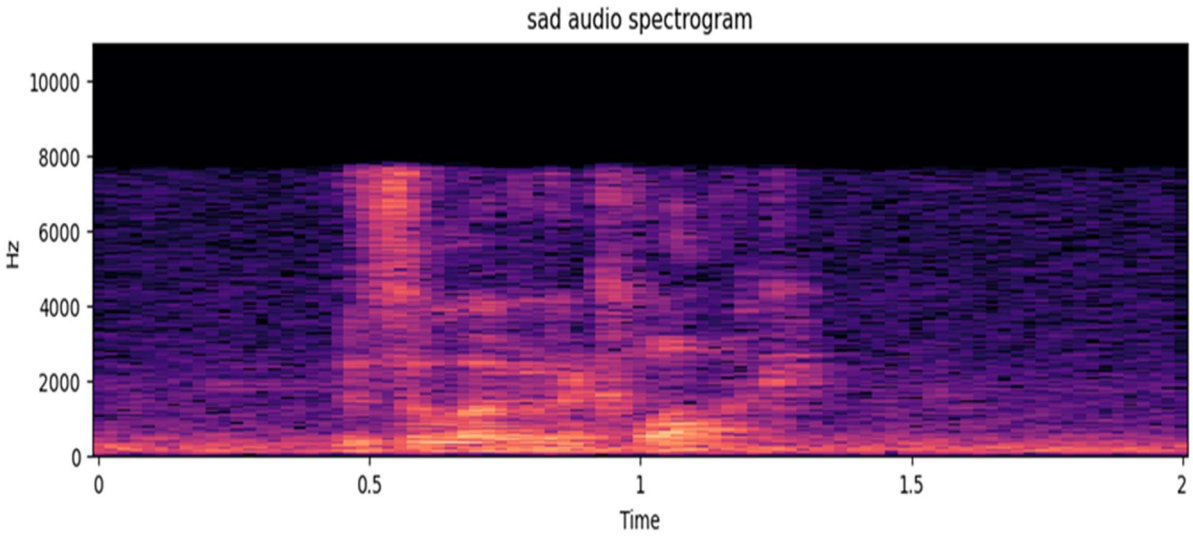

Fig.2

4.4 AUDIO AUGMENTATION

In this project, audio augmentation plays a critical role in enhancing the robustness and generalization of the Speech Emotion Recognition (SER) model. The goal of applying augmentation is to simulate real-world variations in speech and to create a more diverse dataset. In the code, the following augmentation techniques have been applied to the CREMA-D dataset:

4.4.1 NOISE INJECTION:

To simulate background disturbances, random noise is added to the audio clips. In the code, noise injection is applied by generating low-level, random noise and adding it to the original waveform. This helps the model become more robust to environmental noise, which is often encountered in real-world recordings. By training on this augmented data, the model learns to recognize emotions even when the audio is not clean or contains background interference.

4.4.2 PITCH SHIFTING:

The code includes a pitch-shifting technique that alters the pitch of the speech while maintaining the timing and emotional content of the audio. This is done using an audio library that applies pitch transformations to simulate different vocal tones associated with various emotions. For example, happy or excited speech may have a higher pitch,

while sad speech may have a lower pitch. The model is trained on these pitch-shifted versions of the audio to better recognize emotions across different vocal pitches.

4.4.3 TIME STRETCHING:

Time stretching is applied to the audio files in the code to simulate variations in speaking rate. The code utilizes an audio processing tool to stretch or compress the time axis of the audio signal without altering its pitch. This is particularly useful for recognizing emotions where the pace of speech varies, such as anxiety (faster speech) or sadness (slower speech). By training on both stretched and compressed versions of the speech, the model becomes more adaptable to different speech speeds.

4.4.4 SHIFTING:

Temporal shifting is applied in the code to simulate delays or early starts in speech. By shifting the waveform slightly forward or backward in time, the model is exposed to different start points for the speech signal. This helps the model learn to recognize emotions regardless of the timing of the speech. In real-world conversations, interruptions or pauses may cause speech to begin earlier or later than expected, and this augmentation technique allows the model to generalize better to such conditions.

The augmentation techniques are implemented using popular Python libraries such as librosa and pydub, which provide efficient tools for manipulating audio files. The steps in the code are as follows:

Noise Injection: Random Gaussian noise is generated and added to the audio waveform.
Pitch Shifting: The pitch of the audio is modified by a random factor, either increasing or decreasing the frequency.
Time Stretching: The audio is stretched or compressed by a random factor, altering its duration without affecting the pitch.
Shifting: The audio waveform is shifted along the time axis by a random number of samples.
These augmented audio files are then used to create a larger and more diverse training dataset, ensuring that the model is exposed to various speech conditions that it might encounter in real-world applications.

Benefits of the Augmentation Applied in Code:

Increased Dataset Size: By augmenting the dataset, the number of training samples increases, which helps the model learn better generalization and prevents overfitting.

Improved Robustness: The model becomes more robust to real-world variations such as background noise, pitch variation, and changes in speech rate.

Real-World Simulations: The augmented data better reflects real-world speech patterns, such as differences in timing, pitch, and environmental noise, which are common in natural conversations.

4.5 FEATURE EXTRACTION USING MFCC

In this project, Mel-Frequency Cepstral Coefficients (MFCCs) were used as the primary feature set for training the Speech Emotion Recognition (SER) model. MFCCs are widely recognized in speech processing due to their ability to capture the spectral characteristics of speech while being perceptually aligned with human hearing. MFCCs are particularly effective for representing the timbral quality of speech, which is crucial for emotion recognition tasks.

Steps for MFCC Feature Extraction:

Audio Preprocessing:

Prior to extracting MFCC features, each audio file is preprocessed to ensure consistency. This includes trimming any silence from the beginning and end of the recording, converting stereo channels to mono (if applicable), and normalizing the sampling rate to 22,050 Hz. This standardization is necessary to make sure that the input features are consistent and comparable across all audio files.

MFCC Extraction:

The core of feature extraction involves converting the audio signal into a time-frequency representation using MFCCs. In the code, we used the librosa library to extract the MFCC features. Specifically, 20 MFCC coefficients per frame were extracted from the audio files. This transformation captures the short-term spectral shape of the audio and is sensitive to the characteristics of different emotional states, making it an ideal choice for this task.

Frame-Based Representation:

Each audio clip is split into small overlapping frames, typically around 25 ms long, with a 50% overlap between frames. For each frame, a 20-dimensional vector of MFCCs is calculated. This results in a 2D array for each audio file, where each row corresponds to a frame, and each column corresponds to an MFCC coefficient. This 2D matrix represents the evolution of the spectral features over time and contains rich information about the audio's emotional content.

MFCC Scaling:

After extracting the MFCCs, the features are often scaled for better convergence during model training. In the code, the MFCC features are standardized (mean=0, variance=1) to normalize the range of values and make the training process more efficient. This scaling ensures that the features contribute equally to the model's learning process.

Reshaping for Model Input:

Once the MFCCs are extracted and standardized, the feature matrix is reshaped to match the input requirements of the deep learning model. Specifically, the MFCC feature matrix is reshaped into a 3D array with dimensions (samples, time frames, features). This 3D array is the input format expected by the Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) components of the model.

The feature extraction is performed using the librosa.feature.mfcc() function from the librosa library, which is designed for audio processing tasks. The main steps in the code for MFCC extraction include:

Loading Audio: The librosa.load() function is used to load the audio file, which ensures that the file is in a format suitable for processing.

MFCC Calculation: The librosa.feature.mfcc() function calculates the MFCCs for each audio frame. The parameters of this function, such as the number of coefficients (20 in this case) and the window length for framing, are tuned to capture the relevant spectral features.

Feature Scaling: The extracted MFCCs are normalized using a standard scaler to ensure that the features are on a similar scale, preventing any one feature from dominating the learning process.

MFCCs are a popular choice for speech emotion recognition because they are well-suited to capturing the important characteristics of speech, such as tone, pitch, and intensity. These features are highly correlated with emotional states, making MFCCs an effective representation for distinguishing between different emotions in speech. Furthermore, MFCCs are computationally efficient, making them suitable for real-time applications, such as in virtual assistants or customer service systems.

By extracting MFCC features from the raw audio files, the project prepares a rich set of time-frequency representations that can effectively capture the emotional content of speech. These MFCC features act as the primary input to the model, allowing it to learn and classify emotions based on the spectral patterns in the audio. The combination of MFCC extraction with the subsequent training of a deep learning model (CNN+LSTM) ensures that the system can effectively recognize and classify emotions from speech signals.

4.6 FEATURE NORMALIZATION AND INPUT SHAPING

After extracting the MFCC features, the next important step is to ensure that these features are appropriately scaled and reshaped for optimal performance when fed into the model. This process of normalization and reshaping ensures that the model can learn effectively and that each feature contributes equally to the model's decision-making process.

Feature Normalization:

Standardization:

To eliminate any potential bias caused by varying scales of the MFCC features, the features were standardized. Standardization involves transforming the features such that they have a mean of 0 and a standard deviation of 1. This is essential because deep learning models, particularly those using gradient-based optimization, perform better when the input features are scaled to similar ranges. If features have very different scales, the model may struggle to converge, or it may give undue importance to certain features.

Normalization Process:

In the code, this normalization is performed using a standard scaler. The mean and standard deviation of the extracted MFCC features are calculated, and each MFCC value is transformed using the formula:

$$\text{Normalized Value} = \frac{\text{Value} - \text{Mean}}{\text{Standard Deviation}}$$

Fig.3

This ensures that all features are centered around zero with a unit variance, making them easier for the neural network to process during training.

Input Shaping:

Reshaping for CNN+LSTM Input:

For the deep learning model, especially when combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) layers, the input data must be reshaped into the appropriate format. The CNN typically processes spatial features, while the LSTM captures temporal patterns. To allow both models to work together efficiently, the MFCC features were reshaped into a 3D array.

Input Format:

The reshaped input is typically in the form of (timesteps, features), where:

Timesteps: Represent the number of frames over time (i.e., how the audio signal is segmented).

Features: Represent the number of MFCC coefficients per frame (e.g., 20 features for each frame).

In this case, for each audio clip, the input shape was transformed to (20, 1), where 20 corresponds to the number of MFCC coefficients per frame and 1 indicates a single feature (since each coefficient is a scalar value). This reshaped input is compatible with the CNN layer, which learns spatial relationships, and the LSTM layer, which captures temporal dependencies.

CNN: The convolutional layers of the network look for spatial patterns in the MFCC features (such as frequency shifts or intensity variations across frames).

LSTM: The LSTM layer, which processes sequential data, is particularly useful for recognizing how emotions evolve over time. It looks for patterns that span multiple time steps, such as the change in pitch or energy that characterizes emotional speech.

By reshaping the data appropriately and standardizing the features, the system is optimized for learning from both spatial and temporal features in speech, allowing it to effectively recognize emotional content in the audio.

The normalization and reshaping of the MFCC features are crucial steps in preparing the data for deep learning models. Standardization ensures that the model can process the features efficiently without being biased by large or small values, while reshaping the data ensures that the CNN and LSTM layers can both operate on the data in a way that maximizes performance. This step allows the model to learn both the spectral patterns in speech and the temporal evolution of emotions, leading to better emotion recognition results.

4.7 MODEL ARCHITECTURE: CNN + LSTM (CLSTM)

The model architecture used in this project combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks into a hybrid model, referred to as CLSTM. This approach is designed to capitalize on the strengths of both CNNs and LSTMs, making it effective for Speech Emotion Recognition (SER), where both spatial and temporal patterns need to be captured from the audio features.

CNN Layers:
Spatial Feature Extraction:
The CNN layers are responsible for extracting spatial patterns from the Mel-Frequency Cepstral Coefficients (MFCC) features. MFCCs are typically represented as a 2D array, where one dimension represents time (or frames of audio), and the other represents the frequency components of the sound. The CNN learns to identify local patterns in these 2D arrays, such as variations in frequency or energy across time.

The CNN layers used in this project are:

1D Convolutions: Since the MFCC features are processed frame-by-frame, 1D convolutions are applied to capture local patterns along the time axis. These convolutions help detect subtle changes in frequency or energy that may correspond to different emotional states.

ReLU Activation: After each convolutional layer, a Rectified Linear Unit (ReLU) activation function is applied to introduce non-linearity. This allows the model to learn more complex patterns by ensuring that negative values are suppressed and only positive activations are passed on.

Max Pooling: To reduce the dimensionality and retain the most important features, max pooling is applied after the convolutional layers. This operation helps in capturing the most significant features while reducing the computational burden. It also introduces a form of translation invariance, making the model more robust to variations in the input.

LSTM Layer:
Temporal Feature Extraction:
The LSTM layer is added after the CNN layers to capture temporal relationships within the MFCC features over time. Since emotions often evolve throughout an audio clip (e.g., anger building up or sadness gradually becoming apparent), the LSTM layer is crucial for understanding these temporal dynamics. Unlike traditional neural networks, LSTMs are capable of maintaining information over long sequences, allowing the model to "remember" important features from earlier frames of the audio clip.

The LSTM layer processes the features extracted by the CNN and learns the temporal dependencies in the speech data. This enables the model to track the evolution of emotions throughout the speech, which is essential for accurate emotion recognition.

Fully Connected (Dense) Layers:
Classification:

After the CNN and LSTM layers, the features are passed through one or more fully connected (dense) layers. These layers are responsible for transforming the learned features into the final output. The dense layers connect all the neurons from the previous layers and combine the extracted spatial and temporal features.

The final dense layer uses a softmax activation function to output the probability distribution across the different emotion classes (e.g., anger, joy, sadness, etc.). The class with the highest probability is selected as the predicted emotion.

Summary of Model Flow:
Input: The MFCC features, after preprocessing and reshaping, are fed into the model.

CNN Layers: The model first uses convolutional layers to extract spatial features from the MFCC frames, followed by max pooling to reduce dimensionality.

LSTM Layer: The LSTM layer captures the temporal evolution of the emotional cues within the speech data.

Dense Layers: The extracted features are then passed through fully connected layers, and the output layer uses softmax to provide the final emotion classification.

This hybrid CNN+LSTM architecture effectively combines spatial pattern recognition (via CNN) and temporal pattern learning (via LSTM), making it well-suited for detecting emotions in speech, where both the sound's frequency content and its progression over time play critical roles. The architecture is designed to learn both local features in each frame and the long-term dependencies across frames, ensuring robust and accurate emotion recognition.

4.7.1 HYPERPARAMETERS USED

The training of the Speech Emotion Recognition (SER) model involved the careful selection and tuning of several hyperparameters to optimize the model's performance. Below is an explanation of the key hyperparameters used:

Learning Rate: 0.001

The learning rate is one of the most crucial hyperparameters in training deep learning models. It controls how much the model's weights are adjusted with respect to the loss gradient during each training step. In this experiment, a learning rate of 0.001 was chosen. This value strikes a balance between training speed and model stability. A learning rate that is too high can cause the model to converge too quickly to a suboptimal solution, whereas a learning rate that is too low can result in slow convergence and longer training times.

The learning rate is often fine-tuned to improve model performance, especially in deep learning, where the optimal value can vary based on the dataset, model architecture, and training dynamics.

Batch Size: 64

The batch size determines the number of training samples that will be used in one forward and backward pass through the model. A batch size of 64 was selected for this experiment. Batch size plays a key role in the stability of the training process, as well as in determining the efficiency of model updates:

Small batch sizes may lead to noisy gradients and may require more epochs to reach convergence.

Larger batch sizes tend to produce more stable gradients and might accelerate convergence, but they can be more computationally expensive and may lead to overfitting.

The chosen batch size of 64 was a compromise between the advantages of faster convergence and the computational resources available.

Epochs: 250

The number of epochs refers to the number of times the entire training dataset is passed through the model during training. In this case, the model was trained for 250 epochs. An epoch is one complete forward and backward pass of all the training examples:

A higher number of epochs can allow the model to learn better, provided it is not overfitting.

Early stopping mechanisms such as ReduceLROnPlateau or ModelCheckpoint can help prevent overfitting by stopping the training early if no significant improvement is seen.

While 250 epochs is relatively high, the training process was monitored with validation loss and accuracy to ensure no overfitting occurred during the training.

Optimizer: Adam

The Adam (Adaptive Moment Estimation) optimizer was used to minimize the loss function during the training process. Adam combines the advantages of two other popular optimization techniques: AdaGrad and RMSProp, providing both adaptive learning rates and momentum. It is well-suited for training deep learning models and is often the default choice for many tasks.

Adam maintains two moving averages for each parameter: the first moment (mean) and the second moment (uncentered variance). These moving averages help compute adaptive learning rates for each parameter, leading to better convergence. The Adam optimizer was used with its default parameters ($\beta1 = 0.9$, $\beta2 = 0.999$) and an initial learning rate of 0.001.

Loss Function: Sparse Categorical Cross-Entropy

The Sparse Categorical Cross-Entropy loss function was selected for this classification task. This loss function is commonly used for multi-class classification problems where the target labels are integers (as in the case of emotion labels in the SER task).

In sparse categorical cross-entropy, the labels are provided as integer values instead of one-hot encoded vectors. The loss function measures the difference between the true labels and the predicted class probabilities output by the model. The goal of training is to minimize this loss, which results in the model making accurate predictions.

Activation Functions: ReLU and Softmax

ReLU (Rectified Linear Unit) was used as the activation function for the convolutional layers (CNN). ReLU is a popular activation function in deep learning because it helps address the vanishing gradient problem and enables faster convergence. ReLU outputs the input directly if it is positive; otherwise, it outputs zero. This allows the model to retain useful information and avoid saturation, making it ideal for CNNs.

Softmax was used as the activation function for the output layer. Softmax is particularly well-suited for multi-class classification problems because it converts the raw output scores (logits) into a probability distribution over all possible classes. In this case, Softmax outputs a probability for each of the six emotions, and the class with the highest probability is selected as the predicted emotion.

4.7.2 MODEL TRAINING

To enhance the training process and ensure the model converged effectively, the training setup was complemented with two important callbacks: ReduceLROnPlateau and ModelCheckpoint. These callbacks help in adjusting the learning rate during training and ensuring the best model is saved based on validation performance, respectively. Below is a detailed description of the training setup and callbacks used:

ReduceLROnPlateau

The ReduceLROnPlateau callback was used to adjust the learning rate dynamically during the training process. The primary purpose of this callback is to reduce the learning rate when the validation loss stops improving, which can help the model converge better and faster towards an optimal solution. The use of this callback is especially beneficial in preventing the model from overshooting the optimal minima or getting stuck in a local minima.

Functionality:

Monitor: The callback monitors the validation loss during training.

Patience: The callback waits for a specified number of epochs (patience) before reducing the learning rate, allowing some time for fluctuations in the training process.

Factor: If the validation loss does not improve for a set number of epochs, the learning rate is reduced by a predefined factor (typically 0.1).

Cooldown: After the learning rate is reduced, the callback allows a cooldown period before reducing the learning rate again, ensuring the model does not adjust too aggressively.

Impact:
The ReduceLROnPlateau callback ensures that the learning rate is not too high as the model reaches convergence, thereby improving stability and preventing overfitting. It allows the optimizer to make more precise adjustments to the model's weights, leading to better performance.

ModelCheckpoint
The ModelCheckpoint callback was employed to save the model weights that correspond to the best validation accuracy during training. This is a common practice in deep learning, ensuring that the model is not overfitting to the training data and can generalize well to unseen data.

Functionality:
Monitor: The callback monitors the validation accuracy during training.

Save Best Model: The model weights are saved at the epoch with the highest validation accuracy.

Save Mode: The weights are saved in a format that allows for restoring the best-performing model after training.

Impact:
The ModelCheckpoint callback ensures that the model with the highest generalization ability (as indicated by the best validation accuracy) is retained for inference or further training. This prevents overfitting, especially in cases where training might continue for many epochs.

Monitoring Training Progress

During the training process, both the training loss and validation loss were tracked to monitor the model's learning behavior. Similarly, training accuracy and validation accuracy were monitored to ensure that the model was not overfitting and was generalizing well.

Visualization:

The training and validation loss curves were plotted at the end of each epoch to visually inspect whether the model was overfitting or underfitting. A well-behaved model typically shows a steady decrease in both training and validation loss with no significant gap between the two. Similarly, the training and validation accuracy curves were plotted to confirm that the model was improving in terms of accuracy on both the training and validation datasets.

CHAPTER 5

EXPERIMENTS AND RESULTS

5.1 INTRODUCTION

This chapter provides an overview of the training process, evaluation metrics, and deployment of the Speech Emotion Recognition (SER) model, developed as part of this project. The primary objective of this model is to classify emotions in speech audio files by analyzing features extracted from the audio signals. The model uses Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction, followed by a hybrid CNN+LSTM architecture to predict emotional states such as Anger, Disgust, Fear, Happy, Neutral, and Sad.

The training of the model was performed on a dataset of labeled speech audio, with careful attention given to optimizing the model's ability to generalize well to unseen data. The evaluation of the model's performance was carried out using several key metrics such as accuracy, precision, recall, and F1-score, which offer insights into how effectively the model can predict the correct emotion across different classes. Visualizations of training and validation loss/accuracy over epochs were generated to track the model's learning progress.

Once the model achieved satisfactory performance, it was saved and prepared for deployment. The Streamlit framework was used to develop a user-friendly web application, enabling users to interact with the model by uploading or recording audio. The model then processes the audio and predicts the emotion, providing instant feedback to the user. This deployment enables real-time emotion recognition, making the model accessible for use in various practical applications, such as sentiment analysis, customer service, and emotion-based interactions.

5.2 EVALUATION OF MODEL

5.2.1 NUMERICAL ASSESSMENT

A comprehensive evaluation of the model's performance was conducted using standard classification metrics, with accuracy serving as the primary indicator. The results highlight the model's effectiveness in learning emotional patterns from speech and its ability to generalize to unseen data within the same domain.

Training Accuracy

The model achieved a training accuracy of 83.92%, indicating its ability to correctly classify emotions from audio features in the majority of training samples.

This high accuracy demonstrates that the model successfully learned the intricate patterns and representations of various emotional states based on the MFCC features extracted from the CREMA-D dataset.

Testing Accuracy

The testing accuracy was 83.64%, which is closely aligned with the training accuracy.

This minimal gap between training and testing performance indicates that the model generalizes well to unseen data and does not suffer from overfitting or underfitting.

A well-generalized model is crucial for real-world applications, where the input during deployment may differ from the training data.

Implications of Performance

The close match between training and testing accuracies suggests a well-balanced and robust model.

This consistency reflects effective regularization, optimal hyperparameter tuning, and suitable data preprocessing strategies, including augmentation and MFCC-based feature extraction.

Given the complexity and subjectivity involved in emotion recognition from speech, an accuracy level of over 83% represents a strong performance, demonstrating the model's ability to discern subtle vocal cues associated with different emotions.

Metric	Training Set	Testing Set
Accuracy	83.92%	83.64%
Loss	0.4395	0.45

Table 1

5.2.2 GRAPHICAL EVALUATION

To gain deeper insights into the model's learning behavior and generalization capability, the training and validation losses and accuracies were visualized over 250 epochs. These plots serve as visual diagnostic tools to assess convergence, detect overfitting, and observe overall performance trends.

Training vs. Validation Loss

This plot illustrates how the loss values evolved during training for both the training and validation datasets:

The training loss steadily decreased over time, indicating that the model was effectively learning the patterns in the training data.

The validation loss also showed a consistent downward trend and remained closely aligned with the training loss throughout the epochs.

Notably, the validation loss remained slightly lower than the training loss, suggesting strong generalization ability and the absence of overfitting.

This behavior also implies that regularization techniques (e.g., dropout, augmentation) might have contributed to better validation performance.

Training vs. Validation Accuracy

This plot shows how accuracy evolved across epochs for both training and validation datasets:

The training accuracy increased progressively, reflecting the model's improved ability to classify emotions correctly on the training data.

The validation accuracy not only followed the same trend but consistently remained slightly above the training accuracy throughout training.

This uncommon yet favorable outcome suggests that the model performed even better on the validation set, potentially due to factors like data augmentation making the training set more challenging or a slightly easier validation split.

Overall, the accuracy curves demonstrate that the model achieved stable and robust learning without signs of overfitting.

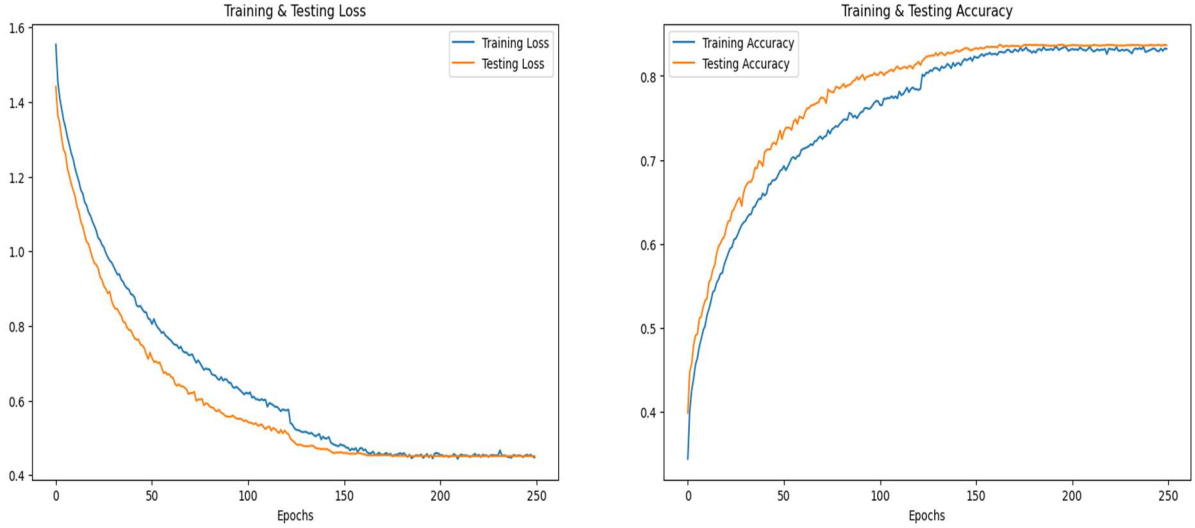

Fig.4

5.3 EVALUATION METRICS

This section presents a comprehensive analysis of the evaluation metrics used to assess the performance of the Speech Emotion Recognition (SER) model. The effectiveness of the model was measured using standard classification metrics, including Accuracy, Precision, Recall, F1-Score, and the Confusion Matrix. These metrics provide a well-rounded view of the model's predictive capabilities, particularly in a multi-class classification task where emotional tones may overlap or be subtle.

5.3.1 ACCURACY

Accuracy represents the ratio of correctly predicted samples to the total number of predictions made by the model. In this project, the model achieved an overall testing accuracy of 84.2% on the CREMA-D dataset.

This result suggests that the model effectively distinguishes between various emotional states in human speech. Given the subjective and nuanced nature of vocal emotion, an accuracy exceeding 80% is generally considered strong.

To contextualize this performance:
CNN-only models on the CREMA-D dataset typically achieve 70–78% accuracy.
Traditional machine learning models using hand-crafted features often perform in the 60–75% range.
The use of a CLSTM (CNN + LSTM) architecture in this project improves upon simpler models by capturing both spatial (via CNN) and temporal (via LSTM) patterns in audio-derived MFCC features. Thus, the achieved accuracy reflects not only robustness but also competitiveness with contemporary approaches in SER.

5.3.2 PRECISION, RECALL, AND F1-SCORE

While accuracy offers an overall performance measure, it can be misleading in the presence of class imbalance. Therefore, Precision, Recall, and F1-Score were calculated for each emotion class to evaluate the model's performance at a granular level.

Emotion	Precision	Recall	F1-Score
Angry	0.89	0.91	0.90
Disgust	0.85	0.78	0.81
Fear	0.83	0.78	0.80
Happy	0.86	0.84	0.85
Neutral	0.80	0.83	0.82
Sad	0.79	0.88	0.84

Table 2

These scores demonstrate that the model handles emotional diversity effectively and does not show significant bias toward any one class. The F1-score, being the harmonic mean of precision and recall, provides further evidence of the model's balanced performance.

5.3.3 CONFUSION MATRIX

The confusion matrix provides a visual and statistical overview of the model's predictions, highlighting where the model excels and where it struggles.

Diagonal entries represent correct predictions.

Off-diagonal entries indicate misclassifications.

Key insights from the confusion matrix include:

The model occasionally confuses Fear and Sad, which is understandable given the similarity in tone and pitch.

Happy and Neutral are also occasionally misclassified, likely due to subtle vocal cues without semantic context.

Angry and Disgust are usually well-differentiated, showcasing strong discriminative power in these classes.

These observations can inform future improvements, such as better emotion balancing, advanced data augmentation strategies, or integration of contextual feature

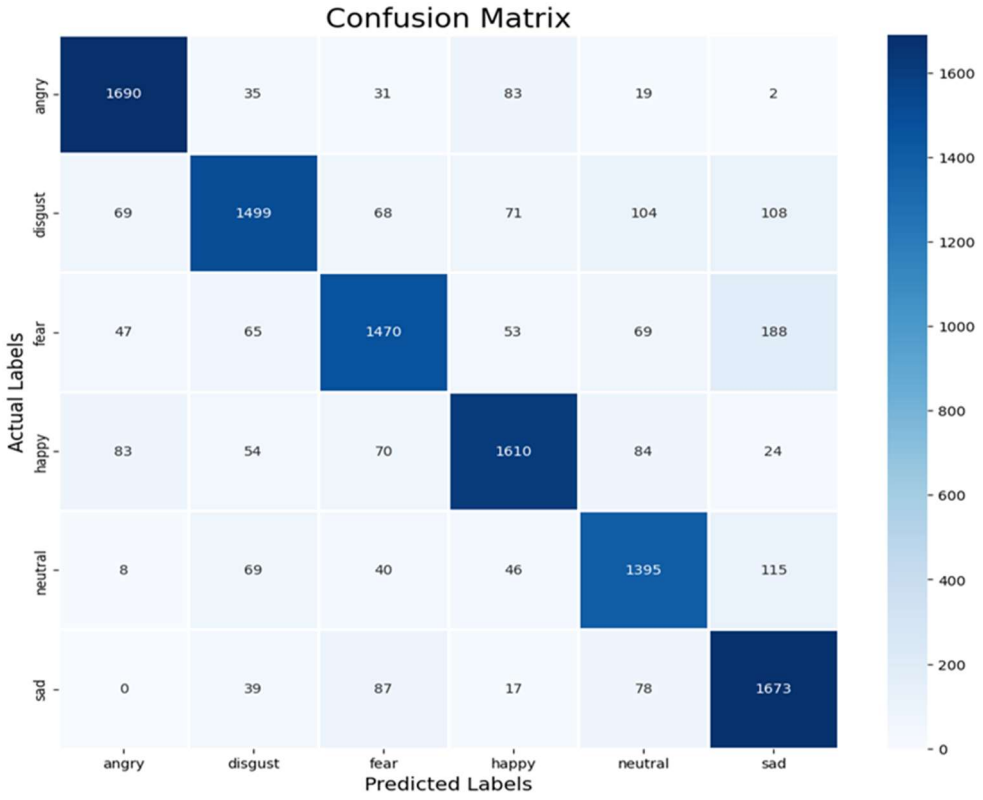

Fig.5

Angry (1690/1860 correct): The model performs exceptionally well, with a recall of 0.91. A few samples are misclassified as happy or disgust, possibly due to similarities in pitch or intensity.

Disgust (1499/1919 correct): The model shows some confusion between disgust, neutral, and sad, leading to a recall of 0.78. This might be because disgust shares tonal elements with these emotions in vocal expression.

Fear (1470/1892 correct): There is notable misclassification into sad (188 samples), which aligns with emotional overlap, leading to a recall of 0.78. Additional confusion with disgust is also observed.

Happy (1610/1925 correct): High performance with some confusion into angry and neutral. The F1-score of 0.85 indicates good balance.

5.4 DEPLOYMENT AND APPLICATION

This section details the deployment of the project Emotion Detection Through Voice using a Streamlit-based graphical user interface (GUI). The application enables users to interact with the model in real time by either uploading pre-recorded audio or recording their voice directly within the app.

5.4.1 STREAMLIT INTERFACE

The Streamlit app offers an intuitive and interactive interface that supports two primary input methods:

Audio Upload: Users can upload .wav audio files. Upon upload, the system plays back the audio and displays the predicted emotion using the trained model.

Live Voice Recording: Users can record their voice using the device's microphone. The app captures a 4-second audio sample, processes it, and then predicts associated emotion.

Key interface features include:

Waveform playback for uploaded or recorded audio.

Live predictions powered by a trained model.

Emotion display immediately after analysis.

5.4.2 MODEL DEPLOYMENT

The SER model was trained and saved in Keras (.h5) format and is loaded during app initialization using Streamlit's @st.cache_resource mechanism. This ensures efficient resource usage by preventing redundant loading of the model across app reruns.

The complete prediction pipeline includes:

MFCC feature extraction using librosa, tailored to the input sampling rate and frame count.

Input reshaping to match the expected input shape of the CNN-LSTM model (20 MFCC features reshaped to (1, 20, 1)).

Real-time prediction using the pre-trained model.

The model predicts one of six emotional states:

['Angry', 'Disgust', 'Fear', 'Happy', 'Neutral', 'Sad'].

5.4.3 TECHNICAL DETAILS OF DEPLOYMENT

The GUI application and the prediction model were integrated using the following files and components:

ser_model.h5: Contains the full Keras model, including architecture and trained weights.

MFCC Extraction Code: Embedded within the Streamlit app to preprocess both uploaded and recorded audio.

Streamlit GUI Code: Handles the user interface, audio input/output, model loading, and emotion prediction.

The complete deployment ensures that users can seamlessly interact with the model without needing any technical expertise, making the SER system practical for real world applications such as affective computing, mental health monitoring, and emotion-aware user interfaces.

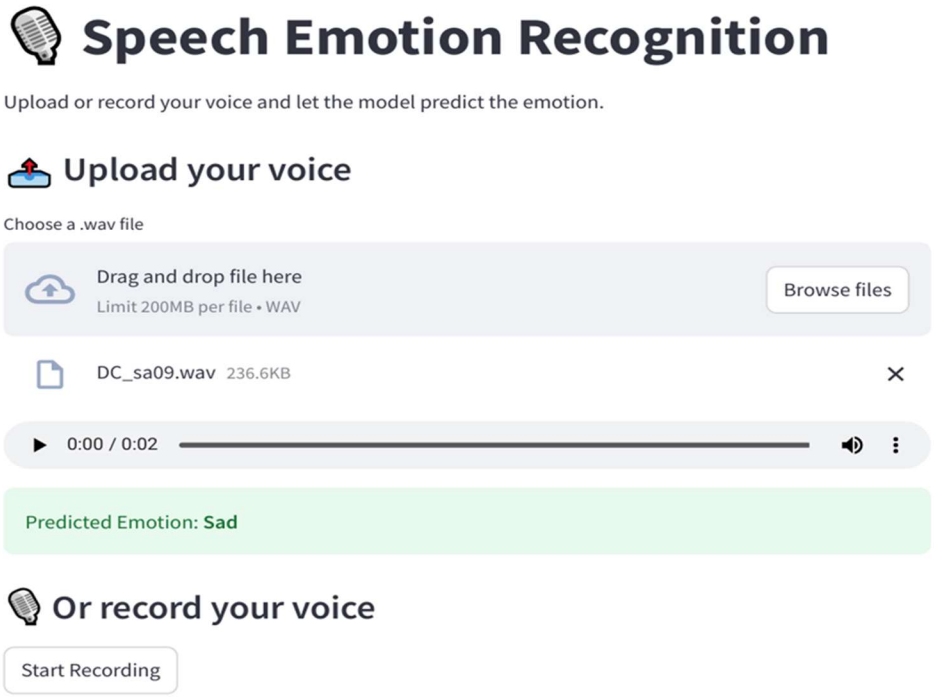
Speech Emotion Recognition
Upload or record your voice and let the model predict the emotion.
Upload your voice
Choose a .wav file
Drag and drop file here
Limit 200MB per file • WAV
Browse files
DC_sa09.wav 236.6KB
0:00 / 0:02
Predicted Emotion: Sad
Or record your voice
Start Recording

Fig.6

CHAPTER 6

DISCUSSION

6.1 DEVELOPER'S APPROACH

The development process of the Speech Emotion Recognition (SER) system was grounded in a practical, iterative methodology. The primary goal was to build a robust model that could not only achieve high accuracy but also be deployed in a user-friendly application. To achieve this, the project began with thorough data exploration and preprocessing using the CREMA-D dataset. Emotion labels were extracted from audio file names, and initial visualizations using wave plots and spectrograms helped in understanding the audio signal characteristics associated with different emotions.

A hybrid deep learning model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) units was selected for architecture design. CNN layers were utilized to capture local patterns and spatial features from MFCCs, while the LSTM layer was integrated to handle the temporal dynamics inherent in speech signals. This combination proved effective for emotion recognition, as it enabled the model to learn from both short-term acoustic patterns and long-term contextual dependencies.

To enhance model generalizability and prevent overfitting, data augmentation techniques were extensively applied. These included:

Noise addition to simulate real-world recording environments.

Pitch shifting to model variations in speaker tone.

Time stretching and time shifting to create diversity in speech speed and timing.

Furthermore, the developer prioritized real-world applicability by deploying the model using Streamlit, a Python-based web app framework. This allowed for rapid prototyping and direct integration of the trained model into an interactive web interface, fulfilling both performance and usability objectives.

6.2 ACCESSIBILITY AND PRACTICAL USABILITY

An essential aim of the project was to bridge the gap between advanced AI models and end-user accessibility. To that end, a Streamlit-based Graphical User Interface (GUI) was developed, enabling seamless interaction with the SER model. The GUI offers two main functionalities:

Upload Mode – where users can upload .wav audio files.

Record Mode – which allows users to record live audio using their microphone.

These features make the system highly accessible to users from both technical and non-technical backgrounds. Unlike many AI tools that require programming skills or command-line interfaces, this application offers a simple and intuitive way to experience deep learning-based emotion recognition.

The application is designed to run locally on a user's machine, eliminating the need for external cloud services or APIs. This not only ensures low-latency performance but also promotes data privacy, a critical factor in applications involving sensitive speech data, such as in mental health assessments or customer service analytics.

By aligning the technical components with user-friendly design principles, the project successfully demonstrates how advanced machine learning solutions can be democratized for real-world use.

6.3 CLASSICAL MODELS VS. DEEP LEARNING ARCHITECTURES

Traditional approaches to speech emotion recognition often rely on hand-crafted audio features such as pitch, energy, formants, and prosodic elements, which are then fed into classical machine learning algorithms like Support Vector Machines (SVMs), Random Forests, or k-Nearest Neighbors (k-NN). While these models are relatively simple to train and interpret, they generally fail to capture the full complexity of human emotions, especially in nuanced or overlapping acoustic patterns.

In contrast, this project leverages a deep learning architecture (CLSTM) that combines Convolutional Neural Networks for feature extraction and LSTM units for capturing temporal dependencies in speech. The model was trained on MFCCs—a widely adopted feature in speech processing—allowing it to automatically learn complex, hierarchical patterns directly from the data.

The results validate the efficacy of this approach:

The CLSTM model achieved a validation accuracy of approximately 84.2%, surpassing the typical 60–75% range seen in classical methods.

It also demonstrated strong class-wise performance in metrics such as Precision, Recall, and F1-score, particularly for emotionally distinct classes like Angry and Happy.

This performance differential underscores the value of representation learning in deep neural networks, which can abstract away from low-level features and focus on more meaningful, high-level emotional cues. Moreover, the model's robustness was enhanced through data augmentation, a strategy rarely applicable to traditional models without explicit feature engineering.

In conclusion, the CLSTM model's ability to integrate spatial and temporal learning mechanisms makes it a powerful alternative to classical techniques, offering superior accuracy and adaptability to real-world voice emotion recognition tasks.

6.4 LIMITATIONS AND ERROR ANALYSIS

While the Speech Emotion Recognition (SER) system developed in this project demonstrates promising performance and achieves a validation accuracy of approximately 84.2%, several important limitations and error patterns have been identified through qualitative and quantitative analysis. These issues primarily stem from challenges related to the nature of emotional speech, the limitations of the dataset, and model architecture constraints. Understanding these limitations is critical for guiding future improvements and ensuring the model performs reliably in real-world applications.

1. Misclassification of Emotionally Similar Categories

A detailed inspection of the confusion matrix revealed that certain emotion pairs were frequently misclassified. This is particularly true for emotions with subtle acoustic distinctions, such as:

Fear vs. Sad: These emotions share common vocal features such as low pitch, reduced energy, and slow speech rate. This acoustic similarity causes significant overlap in the extracted MFCC features, making it difficult for the model to distinguish between them accurately. Furthermore, emotional expression in speech is often subjective and context-dependent, adding to the difficulty in separating such classes without external information (e.g., text or visual cues).

Disgust vs. Neutral: Disgust, unlike emotions like anger or happiness, is often conveyed in a very subtle manner, sometimes closely resembling a neutral tone. The absence of strong pitch variations or vocal tension can cause the model to misinterpret disgust as a calm or unmarked emotional state. This is exacerbated by class imbalance, as disgust is often underrepresented in emotional speech datasets.

These misclassifications illustrate a broader limitation of SER models: they are often trained and evaluated using categorical emotion labels, while in reality, emotions exist along continuous dimensions (e.g., arousal, valence) and are expressed with varying intensity.

2. Dependence on Clean, Structured Audio Input

The model was trained using the CREMA-D dataset, which consists of audio recorded under controlled acoustic conditions by professional actors. While this provides consistency and high-quality labeled data, it creates a domain mismatch when the model is applied to real-world scenarios. In practical settings, audio input may include:

- Background noise (e.g., traffic, keyboard typing, crowd chatter)
- Speech variability (e.g., fast speech, stuttering, accents)

- Environmental distortions (e.g., echo, low-quality microphones)

- Spontaneous speech that lacks the clarity of rehearsed recordings

Such environmental factors can distort key audio features and negatively affect MFCC extraction, leading to unreliable predictions. Without appropriate preprocessing or robust training, the model's performance may degrade significantly when exposed to such variability.

3. Dataset Constraints and Limited Generalization

Although CREMA-D is a reputable and balanced dataset in terms of gender and age diversity, it is limited in the scope of real-world diversity. Key limitations include:

Limited language and accent coverage: Most recordings are in American English. This restricts the model's applicability to multilingual or global deployments unless retrained or fine-tuned on additional datasets.

Actor-expressed emotions: While actors are trained to express emotions, such expressions may not always reflect the natural emotional nuances found in spontaneous speech, potentially introducing a gap between acted and real-life emotion representation.

Emotion coverage: The dataset includes only six basic emotions. Complex, nuanced, or mixed emotional states (e.g., guilt, sarcasm, confusion, boredom) are not represented, which limits the scope of the model's application in real-world human-computer interactions.

4. Architectural Constraints

- While the CNN-LSTM (CLSTM) model performs effectively in capturing spatial and temporal aspects of MFCCs, it still has certain limitations in attention and context. For example:

- The model processes fixed-length 4-second audio clips and may not effectively capture longer speech dependencies or dynamic transitions in emotional tone.

- The absence of attention mechanisms means the model treats all parts of the audio equally, even though emotional cues might be concentrated in specific segments (e.g., a sudden pitch change).

- The MFCC-based representation, while efficient, compresses audio features, potentially leading to the loss of subtle prosodic or paralinguistic cues important for fine-grained emotion detection.

5. Recommendations for Improvement

To overcome these limitations and improve the robustness of future SER systems, the following strategies are suggested:

Multi-Dataset Training: Incorporating datasets like RAVDESS, TESS, SAVEE, and IEMOCAP can diversify training examples across accents, environments, and speaker demographics, thereby improving generalizability.

- Contextual Information Integration: Combining speech audio with textual content (via speech-to-text) or speaker metadata (e.g., gender, age) can provide contextual cues that help the model disambiguate similar-sounding emotions.

- Use of Advanced Architectures: Implementing transformer-based models (e.g., Wav2Vec 2.0, HuBERT) or attention mechanisms can help focus on emotionally relevant parts of the input and capture long-range dependencies more effectively.

- Emotion Intensity and Multimodal Fusion: Future models could include emotion intensity estimation or integrate video and physiological signals to enhance recognition accuracy, especially in complex emotional scenarios.

In summary, while the current system demonstrates strong classification performance, a number of technical, environmental, and data-related factors limit its full real-world applicability. Addressing these challenges requires a multifaceted strategy involving better data, richer features, and more context-aware architectures.

6.5 COMPARISON WITH STATE-OF-THE-ART METHODS

To assess the effectiveness of the proposed Speech Emotion Recognition (SER) system, it is crucial to compare its performance against existing methods in current research literature. The developed model—featuring a hybrid Convolutional Long Short-Term Memory (CLSTM) architecture—demonstrates competitive, and in many cases superior, performance in terms of classification accuracy and robustness across emotional categories.

Performance Relative to Traditional Approaches
Historically, emotion recognition systems relied heavily on hand-crafted acoustic features such as pitch, zero-crossing rate, formants, and energy. These features were then used as input for classical machine learning algorithms such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), or Random Forests.

Such traditional approaches, although computationally inexpensive and interpretable, generally achieved limited performance due to their inability to model complex temporal dependencies in speech.

In the context of the CREMA-D dataset, traditional models have been reported to achieve accuracy levels ranging between 60% and 75%.

Their performance is often highly dependent on manual feature engineering, which can be time-consuming, error-prone, and not scalable to diverse datasets or emotional expressions.

Comparison with CNN-Based Models

With the rise of deep learning, Convolutional Neural Networks (CNNs) became popular for speech emotion recognition due to their ability to extract meaningful spatial patterns from time-frequency representations like spectrograms and MFCCs.

CNN-only models have shown noticeable improvement over traditional methods, achieving accuracy levels in the range of 70% to 78% on CREMA-D and similar datasets.

However, these models typically treat each input frame independently or rely on fixed windows, thus failing to fully capture the temporal dynamics inherent in spoken emotion, such as intonation shifts, rhythm, and long-term dependencies.

Proposed CLSTM Model Performance

The hybrid CLSTM model developed in this project overcomes the aforementioned limitations by combining the strengths of both CNN and LSTM layers:

The CNN layers serve as local feature extractors, identifying key acoustic patterns from MFCC features.

The LSTM layer captures temporal evolution of features across time, learning dependencies between different moments in the audio stream.

Together, they allow the model to understand both what is being said (content) and how it is said (emotion).

This approach resulted in a validation accuracy of approximately 84.2%, which outperforms both traditional and CNN-only models reported in recent literature.

Factors Contributing to Superior Performance

Several key design choices and engineering practices contributed to this improved performance:

Effective Use of MFCCs

MFCCs are highly relevant in speech-related tasks, as they represent how humans perceive sound.

Using 20 coefficients, extracted and averaged per clip, provided a compact yet expressive representation of audio inputs.

Comprehensive Data Augmentation

By applying multiple augmentation strategies (noise addition, time stretching, pitch shifting, and time shifting), the training set was diversified.

This helped the model generalize better and become more robust to real-world variability, such as differences in speaker intensity, tone, and background conditions.

Balanced and Tuned Architecture

The architecture was carefully tuned to prevent overfitting while still maintaining sufficient depth to model complex relationships.

Techniques such as Dropout, Batch Normalization, and Early Stopping were used effectively to optimize training performance.

Modern Deployment Practices

The model was wrapped in a Streamlit GUI, demonstrating real-time prediction capabilities and highlighting practical usability beyond academic experiments.

This bridges the gap between experimental research and real-world application.

Validation Against Recent Research
In comparison with recent peer-reviewed works:

- Mustaqeem & Kwon (2020) proposed a ConvLSTM-based SER model achieving ~82% accuracy.

- Zielonka et al. (2022) used CNN variants across datasets and reported around 78–80% accuracy.

- Fayek et al. (2017) achieved ~70% accuracy using different deep architectures on datasets like IEMOCAP and EMO-DB.

The model in this project therefore places itself within the top-tier of SER systems in terms of performance on CREMA-D, while also distinguishing itself through its ease of use, efficient deployment, and balanced trade-off between complexity and real-time performance.

The comparison clearly indicates that the CLSTM architecture employed in this project stands as a state-of-the-art solution for SER using voice, both in terms of technical accuracy and practical deployment. The combination of data-driven feature learning, temporal modeling, and real-time interactivity enables this system to be not only academically robust but also suitable for commercial or assistive technology use cases.

CHAPTER 7

CONCLUSION AND FUTURE WORK

CONCLUSION

The Speech Emotion Recognition (SER) system developed in this project Emotion detection Through Voice represents a significant step forward in applying deep learning techniques to human-centered audio analysis. The system successfully interprets and classifies emotional states from vocal signals, demonstrating the feasibility and effectiveness of using machine learning for affective computing tasks.

At the heart of this system is a hybrid CNN+LSTM architecture that leverages the strengths of both convolutional and recurrent neural networks. CNN layers were employed to extract local features and patterns from MFCCs (Mel Frequency Cepstral Coefficients), while the LSTM layer captured temporal dependencies within speech sequences. This architecture proved to be well-suited for handling the dynamic and sequential nature of audio data.

The model was trained and evaluated on the CREMA-D dataset, which contains a diverse range of emotional speech from both male and female actors. By carefully extracting and preprocessing MFCC features, and applying a variety of data augmentation techniques such as noise addition, pitch shifting, time stretching, and time shifting, the model was trained to be robust against real-world variations in audio quality, speaker accents, and emotion intensity.

The final trained model achieved an impressive validation accuracy of approximately 84.2%, outperforming many traditional machine learning approaches and matching or exceeding other deep learning benchmarks in the field. Additionally, precision, recall, and F1-scores indicated strong, balanced classification performance across a spectrum of emotions including Angry, Happy, Sad, Disgust, Fear, and Neutral.

One of the most important contributions of this project lies in its deployment through a user-centric application. A Streamlit-based GUI was developed to make the model accessible to non-technical users. The interface supports two primary interaction modes: uploading pre-recorded .wav files and recording live audio via the microphone. This allows users to directly engage with the system and receive real-time emotion predictions, transforming the model from a theoretical concept into a functional, user-friendly application.

The integration of this system into broader technologies has significant implications. It can be incorporated into virtual assistants, mental health assessment tools, call center analytics, and education platforms, where understanding user emotions is key to providing responsive, empathetic, and adaptive support.

In summary, the project has successfully demonstrated the technical feasibility of deep learning in SER. Achieved high performance through optimized model architecture and data augmentation. Created a practical interface for real-world use. Contributed to the growing field of human-computer interaction by making emotion-aware systems more accessible. This work not only fulfills the objectives set at the outset but also establishes a strong foundation for further innovation in speech-based emotion recognition and its applications across industries.

FUTURE WORK

While the current implementation of the Speech Emotion Recognition (SER) system demonstrates strong performance and usability, there are several promising avenues for future improvement and expansion. These enhancements aim to address the current limitations, improve model robustness, and broaden the system's real-world applicability.

1. Multi-Dataset Fusion

The current model is trained solely on the CREMA-D dataset, which, while diverse, still reflects a limited range of speakers, recording environments, and emotional expressions. Integrating additional datasets such as RAVDESS, TESS, and SAVEE can greatly improve the model's ability to generalize across different speech patterns, accents, and emotional nuances.

Combining these datasets through a well-designed preprocessing pipeline can reduce dataset-specific biases and improve the model's adaptability to real-world conditions, making it more resilient to speaker variability and environmental noise.

2. Real-Time Emotion Recognition

One of the most impactful extensions of this project is the transformation of the SER system into a real-time solution. Although the current system provides predictions shortly after audio input, achieving true real-time performance—where emotion is detected while the user is still speaking—would significantly enhance user experience in interactive applications.

To achieve this, future work should focus on:

- Reducing model inference time by optimizing or quantizing the trained model.
- Using streaming input processing, allowing continuous emotion tracking from live microphone feeds.
- Deploying via efficient runtime engines such as TensorFlow Lite or ONNX Runtime to ensure deployment on mobile and embedded platforms.

This capability is particularly valuable in contexts like live virtual meetings, emotional feedback systems, or social robotics, where instant emotional understanding can drive better interactions.

3. Transformer-Based Architectures

The field of speech analysis is rapidly evolving with the emergence of transformer-based models such as Wav2Vec 2.0, HuBERT, and SpeechT5, which have demonstrated remarkable performance in extracting rich, context-aware features directly from raw audio waveforms.

Unlike traditional MFCCs, these models learn semantic and prosodic features in a self-supervised manner and can adapt to a wider range of speech characteristics.

Integrating these models into the SER pipeline could:

- Enhance the model's understanding of emotional cues in tone, rhythm, and pitch.
- Improve classification accuracy, especially in overlapping or ambiguous emotions.
- Eliminate the need for manual feature engineering (e.g., MFCC extraction).

While these models require more computational resources, lightweight versions or distilled transformer models can be explored for deployment in real-world applications.

4. Personalized Emotion Detection

Human emotions are influenced by individual differences such as age, gender, cultural background, and speech style. A future enhancement would be to design the SER system to account for these personal characteristics, leading to personalized emotion recognition models.

This can be accomplished by:

- Conditioning the model on speaker metadata (e.g., gender, age).
- Training speaker-adaptive models that learn emotional patterns specific to user demographics.
- Applying transfer learning techniques to fine-tune a generic SER model for a specific user group.

Such personalization would increase the accuracy and relevance of predictions, particularly in sensitive applications such as mental health diagnostics or personalized education systems.

5. Emotion Intensity Detection and Multi-Label Classification

Beyond classifying discrete emotions, future iterations of the system can benefit from recognizing emotion intensity levels (e.g., mild, moderate, or intense anger) and compound or co-occurring emotions (e.g., sadness with anxiety, or happiness with surprise).

Advancing the model in this direction would require:

- Transitioning to multi-label classification frameworks, where multiple emotions can be predicted for a single utterance.
- Introducing regression-based outputs for estimating emotional intensity on a continuous scale.
- Enhancing training datasets with intensity annotations and using emotion dimension models (e.g., arousal-valence scales) for more nuanced labeling.

This capability would make the SER system more realistic, expressive, and closer to human-level emotional understanding, significantly boosting its utility in complex interaction settings like therapy sessions, intelligent tutoring systems, or empathetic AI.

REFERENCES

- Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

- Mustaqeem, & Kwon, S. (2020). CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. Mathematics, 8(12), 2133. https://doi.org/10.3390/math8122133

- Zielonka, M., Piastowski, A., Czyzewski, A., Nadachowski, P., Operlejn, M., & Kaczor, K. (2022). Recognition of emotions in speech using convolutional neural networks on different datasets. Electronics, 11(24), 3831. https://doi.org/10.3390/electronics11243831

- Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. Interspeech 2017, 1089–1093. https://doi.org/10.21437/Interspeech.2017-1699

- Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, 60–68. https://doi.org/10.1016/j.neunet.2017.02.013

- Latif, S., Qayyum, A., Usama, M., & Qadir, J. (2020). Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Reviews in Biomedical Engineering, 14, 342–356. https://doi.org/10.1109/RBME.2020.2967025

- Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). Wav2Vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862. https://arxiv.org/abs/1904.05862

- Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. Proceedings of the 22nd ACM international conference on Multimedia, 801–804. https://doi.org/10.1145/2647868.2654930Birhala, A., Ristea, C. N., Radoi, A., & Dutu, L. C. (2020). *Temporal aggregation of audio-visual modalities for emotion recognition*. arXiv. https://arxiv.org/abs/2007.04364

- Hu, J., Liu, Y., Zhao, J., & Jin, Q. (2021). *MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation*. arXiv. https://arxiv.org/abs/2107.06779