A PROJECT REPORT

on

"Speech Emotion Recognition using Deep Learning"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

**BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY**

BY

ARYAN KUMAR	22053928
SHRESHT SONI	22051626
JAYA KISHAN PATEL	22053958

UNDER THE GUIDANCE OF

Mr. Sunil Kumar Sawant

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

April 2025

A PROJECT REPORT

on

"Speech Emotion Recognition using Deep Learning"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

GROUP MEMBER A	ROLL NUMBER A
GROUP MEMBER B	ROLL NUMBER B
GROUP MEMBER C	ROLL NUMBER C
GROUP MEMBER D	ROLL NUMBER D

UNDER THE GUIDANCE OF
GUIDE NAME

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAE, ODISHA -751024
April 2025

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“Speech Emotion Recognition using Deep Learning “

submitted by

ARYAN KUMAR	22053928
SHRESHT SONI	22051626
JAYA KISHAN PATEL	22053958

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024-2025, under our guidance.

Date: 10 /04/2025

Mr. Sunil Kumar Sawant
Project Guide

Acknowledgements

I would like to begin by expressing my heartfelt gratitude to my institute, **Kalinga Institute of Industrial Technology, Bhubaneswar**, for providing an environment that fosters academic excellence and innovation. The institute’s unwavering support and resources have been instrumental in helping me meet the requirements of my project.

I am grateful to all the professors of the Department of Computer Science and Engineering and other departments for their guidance and encouragement.

A special thanks to our project guide, **Mr. SUNIL KUMAR SAWANT** and his teaching assistants for their constant support and valuable feedback throughout my minor project.

Lastly, I am deeply grateful to my family and friends for their unwavering love, support, and encouragement during this journey.

ARYAN KUMAR
SHRESHT SONI
JAYA KISHAN PATEL

ABSTRACT

Speech Emotion Recognition (SER) is a crucial domain in computer and human interaction, focused on improving machine comprehension of human emotions using audio signals. This project utilizes the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC dataset), which contains Hindi audio samples with 8 emotional variations. The approach involves pre-processing audio data, extracting key acoustic characteristics - Mel-frequency cepstral coefficients (MFCC), Mel spectrum, Chroma and Zero crossing rate (ZCR) - and training machine learning (ML) models for emotion classification. These features capture both the spectral and temporal characteristics of speech, enabling effective emotion recognition. Various classification techniques, including deep learning models (DL), are evaluated to improve accuracy. The study also addresses challenges such as speaker variability and data imbalance. The results contribute to SER applications in emotion analysis, virtual assistants, and mental health assessment, offering information for future advances in speech-based emotion recognition.

Keywords: Speech Emotion Recognition (SER); IITKGP-SEHSC; MFCC; Mel Spectrogram; Chroma; Zero-Crossing Rate (ZCR); Machine Learning; Deep Learning; Audio Processing; Emotion Classification.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

SER is a challenging field that aims to identify human emotions from speech signals. Unlike text-based sentiment analysis, it leverages vocal features such as pitch, tone, and rhythm to recognize emotions. The complexity of human emotions, speaker variability, and environmental noise make this task difficult.

The *IITKGP-SEHSC* dataset provides a rich collection of Hindi speech samples, making it a valuable resource for studying emotion classification in an Indian linguistic context [1]. However, the effectiveness of SER models depends on feature extraction and classification techniques. In this project, we utilize **Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrogram, Chroma, and Zero-Crossing Rate (ZCR)** to capture emotion-relevant characteristics, with the aim of improving recognition accuracy.

1.2 Motivation

With the increasing adoption of artificial intelligence in daily life, emotion-aware systems are becoming essential for human-computer interaction. Virtual assistants, call center automation, and mental health monitoring are some applications that benefit from SER. In healthcare, analyzing speech patterns can help detect stress, depression, or anxiety.

Existing SER research focuses primarily on English data sets, but emotions vary between languages and cultures. The *IITKGP-SEHSC* dataset enables a more inclusive study of SER, ensuring improved accuracy for non-English speakers. This project contributes towards advancing SER models tailored to diverse linguistic backgrounds.

1.3 Objective

The main objectives of this project are:

- Preprocess speech data, including noise reduction and normalization.
- Extract key acoustic features: **MFCC, Mel Spectrogram, Chroma, and Zero-Crossing Rate (ZCR)**.
- Train and evaluate deep learning models for emotion classification.
- Address challenges such as speaker variability and dataset imbalance.
- Analyze performance using accuracy, precision, recall, and F1-score.

CHAPTER 2

LITERATURE REVIEW

2.1 Traditional Approaches

SER has evolved over the years, and early studies rely on traditional machine learning (ML) techniques that involve manual feature extraction. These methods primarily focus on prosodic, spectral, and voice quality features to identify emotional patterns from speech signals.

2.1.1 Feature Extraction Techniques

Traditional approaches to SER rely on extracting hand-engineered features that capture emotional cues in speech. Commonly used features include:

- **MFCC features** – Capture the short-term power spectrum of speech and are widely used for speech processing [1, 2].
- **Pitch and Intensity Features** – Represent variations in vocal frequency and loudness, which are critical indicators of emotions [3].
- **ZCR** – Calculates the frequency of sign changes in a speech signal. [4].
- **Chroma Features** – Capture the harmonic content of speech, aiding in tonal analysis [2].
- **HuBert Features** – Self supervised BERT model used to extract features. [4].

2.1.2 Machine Learning Classifiers

Once features are extracted, traditional classifiers are used to map speech data to specific emotional states. Some widely used classifiers include:

- **Support Vector Machines (SVMs)** – Well-suited for managing high-dimensional data and widely applied in SER. [3].

- **Hidden Markov Models (HMMs)** – Model speech as a sequence of states, capturing temporal dependencies [2].
- **Gaussian Mixture Models (GMMs)** – Probabilistic models that estimate the distribution of features across different emotional classes [1].
- **Random Forests and Decision Trees** – Used for classification due to their ability to handle non-linear patterns [2].

While these traditional methods achieved reasonable performance, they struggled with generalization due to variations in speech, speaker dependency, and background noise. The reliance on handcrafted features also made them less adaptable to complex emotional expressions.

2.2 Deep Learning Approaches

With advancements in deep learning, SER has shifted towards data-driven feature extraction and hierarchical representation learning. DL models autonomously extract discriminative features from raw speech signals and spectrogram representations, significantly improving recognition performance.

2.2.1 Neural Network Architectures for SER

Deep learning methods have demonstrated superior accuracy in SER by leveraging large-scale datasets and complex architectures:

- **Convolutional Neural Networks (CNNs)** – Uses spectrograms to extract spatial features, by capturing frequency-based emotional cues [4].
- **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks** – Model temporal dependencies in speech, making them effective for capturing sequential emotion variations [4].
- **Hybrid CNN-RNN Models** – Combine spatial and temporal learning to improve feature representation in SER [4].
- **Transformers and Self-Supervised Models** – Architectures like Wav2Vec2, HuBert allow direct feature learning from raw audio, eliminating the need for handcrafted features [4].

2.2.2 Recent Developments in SER

The adoption of deep learning has led to several breakthroughs in SER:

- **End-to-End Learning** – Modern models process raw waveforms directly, removing the dependency on feature extraction [4].
- **Attention Mechanisms** – Improve emotion recognition by focusing on relevant speech segments [4].
- **Pretrained Speech Models** – Large-scale transformer models trained on diverse datasets improve SER generalization [4].
- **Cross-Lingual Emotion Recognition** – Expanding SER research to non-English datasets, such as *IITKGP-SEHSC*, to improve inclusivity [1].

Basic Concepts

This chapter outlines the fundamental techniques and models used in SER and reviews previous work in the field.

2.1 Feature Extraction Techniques

- **MFCC (Mel-Frequency Cepstral Coefficients)**: Widely used in speech recognition. It models human hearing by capturing the power spectrum of audio signals.
- **Mel Spectrogram**: A time-frequency representation of sound using a Mel scale which helps analyze emotion-relevant frequency components.
- **Chroma Features**: These represent the 12 pitch classes of music and speech. Useful in capturing harmonic and pitch-related emotion cues.
- **Zero-Crossing Rate (ZCR)**: Measures the rate at which the signal changes sign. High ZCR often corresponds to fricatives or emotional excitement.
- **HuBERT Features**: Derived from a pre-trained self-supervised model, useful for fine-grained speech emotion analysis.

2.2 Classifiers and Models

Traditional machine learning models:

- **Support Vector Machines (SVMs)**: Efficient in high-dimensional spaces.
- **Hidden Markov Models (HMMs)**: Good for modeling sequential data.
- **Gaussian Mixture Models (GMMs)**: Used to model probabilistic distributions.
- **Random Forests/Decision Trees**: Handle nonlinear relationships in data.

Deep learning models:

- **Convolutional Neural Networks (CNNs)**: Used to process spectrograms and capture spatial features.
- **Recurrent Neural Networks (RNNs)/LSTM**: Capture sequential dependencies and temporal patterns.
- **CNN-LSTM Hybrid**: Combines the advantages of CNN (feature extraction) and LSTM (sequence modeling).
- **Transformers**: Models like Wav2Vec2 and HuBERT are pre-trained on large audio corpora and used in SER tasks.

Chapter 3

Problem Statement / Requirement Specifications

The project aims to develop a robust system to classify emotions from Hindi speech samples. The system should be able to preprocess raw audio, extract key features, train deep learning models, and evaluate classification performance.

3.1 Project Planning

- Data Collection (IITKGP-SEHSC dataset)
- Preprocessing: Resampling, silence removal, normalization
- Feature Extraction: MFCC, Mel Spectrogram, Chroma, ZCR
- Model Development: CNN, LSTM, CNN-LSTM
- Evaluation: Accuracy, precision, recall, F1-score

3.2 Project Analysis

- Emotional overlap across classes
- High inter-speaker variability
- Imbalance in class distribution
- Real-time applicability and generalization

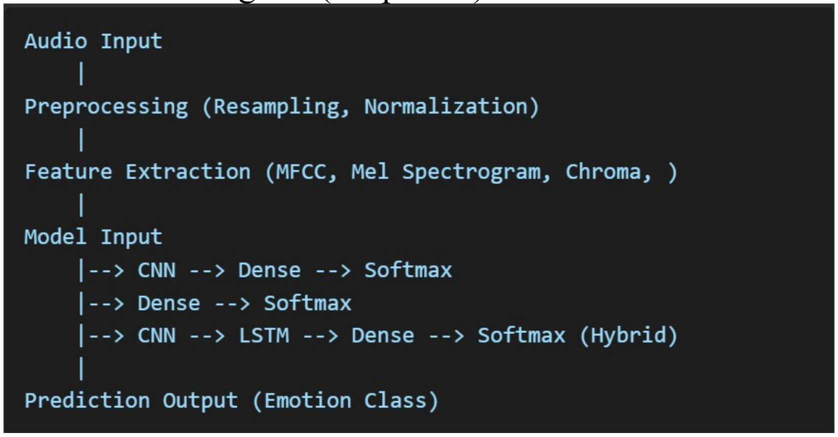
3.3 System Design

3.3.1 Design Constraints

- Software: Python, TensorFlow, Librosa, Keras
- Hardware: GPU-enabled system for faster training

3.3.2 System Architecture

Architecture Diagram (simplified):



```
Audio Input
    |
Preprocessing (Resampling, Normalization)
    |
Feature Extraction (MFCC, Mel Spectrogram, Chroma, )
    |
Model Input
    |--> CNN --> Dense --> Softmax
    |--> Dense --> Softmax
    |--> CNN --> LSTM --> Dense --> Softmax (Hybrid)
    |
Prediction Output (Emotion Class)
```

Chapter 4

Implementation

METHODOLOGY

The workflow of the SER system is illustrated in the following flowchart. It outlines key stages, including pre-processing, feature extraction, model training, and evaluation. Figure 3.1 provides a visual representation of the entire process.

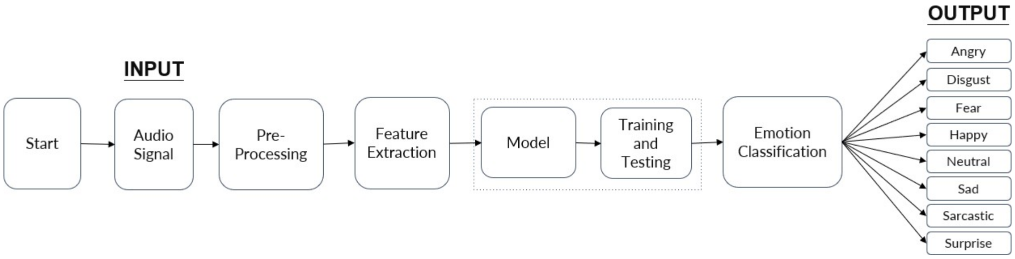


Figure 3.1: Flowchart

3.1 Preprocessing

Preprocessing is a critical step to enhance data quality and optimize feature extraction. The following preprocessing techniques were implemented:

- **Resampling:** All audio signals were resampled to a uniform sampling rate to ensure consistency.
- **Normalization:** Standardizing amplitude variations across samples.

3.1 Feature Extraction Techniques

The following acoustic features were extracted to represent the emotional content of speech effectively:

- **MFCCs):** Extracted from the speech spectrum, capturing short-term power distribution.
- **Mel Spectrogram:** Represents frequency variations over time, providing a detailed spectral analysis.
- **Chroma Features:** Capture tonal characteristics of speech based on the harmonic structure.
- **ZCR):** Measures the rate of signal sign changes, indicative of speech texture.

3.2 Deep Learning Models

For emotion classification, different DL architectures were implemented and evaluated:

3.2.1 Convolutional Neural Networks (CNN)

1D CNN layer architecture was implemented, consisting of:

- Multiple CNN layers utilizing the ReLU activation function for efficient feature extraction.
- Utilized max-pooling layers for dimensionality reduction and retain essential features.
- Fully connected layer to map extracted features to emotion classes.

3.2.2 Recurrent Neural Networks (RNN) and LSTMs

Long Short Term Memory (LSTM) based RNN were utilized to capture temporal dependencies within speech data. The implemented architecture includes:

- LSTM layers to capture sequential dependencies in extracted features.
- Dropout regularization used to prevent overfitting.
- A final dense layer with softmax activation for emotion classification.

3.2.3 Hybrid CNN-LSTM Model

A hybrid model combining CNN and LSTM was developed to leverage both spatial and temporal features:

- CNN layers for extracting spatial features from spectrograms.
- Used LSTM layers to capture long-term dependencies in the extracted features.
- Dense layers for final emotion classification.

3.3 Performance Evaluation

The models were evaluated using standard classification metrics:

- **Accuracy:** Measures the overall correctness of predictions.
- **Precision:** Evaluates the proportion of correctly predicted positive samples.
- **Recall:** Measures the model's ability to identify true positive emotions.
- **F1-Score:** A harmonic mean of precision and recall for balanced evaluation.
- **Confusion Matrix:** Provides a detailed analysis of classification errors.

RESULTS AND INFERENCE

4.1 Experimental Results

The outcomes derived from the SER system are analyzed using various performance metrics The models were trained using the extracted features and tested on a separate validation dataset to measure their effectiveness.

4.1.1 Training and Validation Curves

The training and validation curves of accuracy and loss are illustrated in Figure 4.1, showing model convergence and generalization.

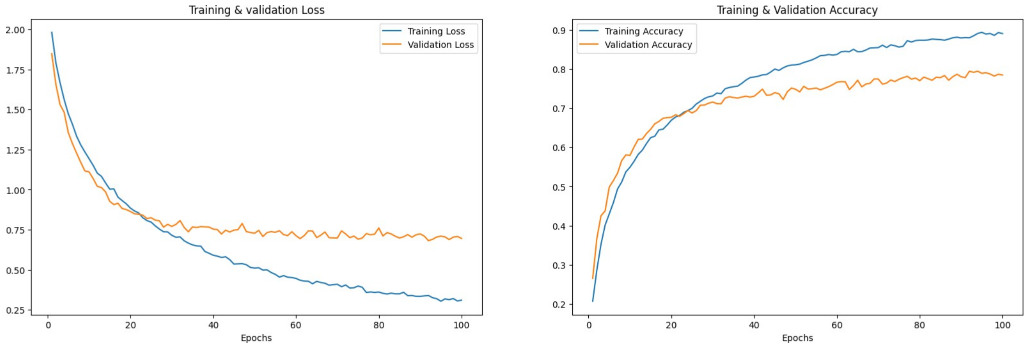
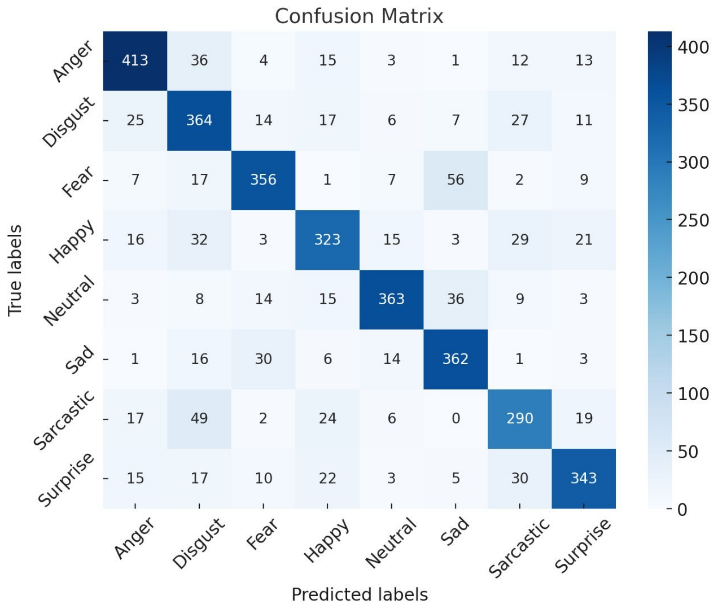


Figure 4.1: Training and Validation Loss/Accuracy Curves

4.1.2 Confusion Matrix

Confusion matrix illustrated in Figure 4.2, which visualizes the model's predictions against true labels.

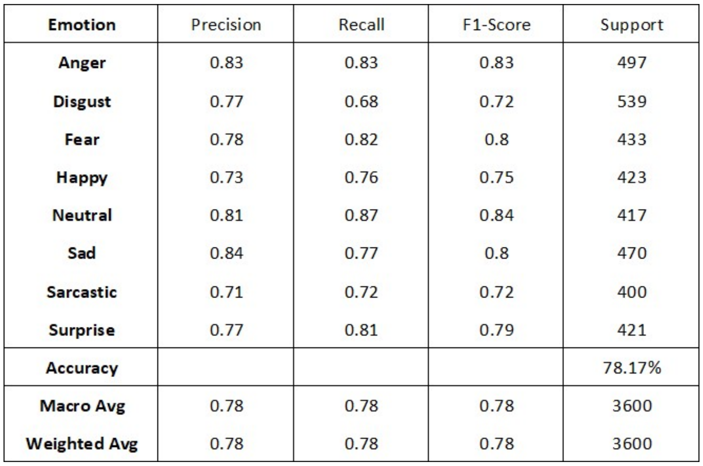


	Anger	Disgust	Fear	Happy	Neutral	Sad	Sarcastic	Surprise
Anger	413	36	4	15	3	1	12	13
Disgust	25	364	14	17	6	7	27	11
Fear	7	17	356	1	7	56	2	9
Happy	16	32	3	323	15	3	29	21
Neutral	3	8	14	15	363	36	9	3
Sad	1	16	30	6	14	362	1	3
Sarcastic	17	49	2	24	6	0	290	19
Surprise	15	17	10	22	3	5	30	343

Figure 4.2: Confusion Matrix of SER Model

4.1.3 Classification Report

The classification report of the best-performing model is presented below in the Figure 4.3, showing precision, recall, and F1-score for every emotion classes.



Emotion	Precision	Recall	F1-Score	Support
Anger	0.83	0.83	0.83	497
Disgust	0.77	0.68	0.72	539
Fear	0.78	0.82	0.8	433
Happy	0.73	0.76	0.75	423
Neutral	0.81	0.87	0.84	417
Sad	0.84	0.77	0.8	470
Sarcastic	0.71	0.72	0.72	400
Surprise	0.77	0.81	0.79	421
Accuracy				78.17%
Macro Avg	0.78	0.78	0.78	3600
Weighted Avg	0.78	0.78	0.78	3600

Figure 4.3: Classification Report of SER Model

4.2 Inference

The experimental results indicate that deep learning models outperform traditional ML classifiers in Speech Emotion Recognition. The final model achieved an accuracy of **78.17%**, demonstrating reliable performance in emotion classification.

4.2.1 Challenges and Limitations

Despite achieving high accuracy, the following challenges persist:

- **Emotional Overlap:** Emotions are not always distinct and can overlap. For example, a person may express a mix of happiness and surprise.
- **Speaker Variability:** The same emotion can be expressed differently by different people, making it hard to create a generalized model.
- **Feature Selection:** Identifying and extracting relevant features that effectively represent emotions is crucial. Overly simplistic features may miss important nuances, while too many features can lead to overfitting.
- **Computational Efficiency:** Balancing the trade-off between capturing enough information and maintaining computational efficiency is challenging.
- **Generalization to Real-World Data:** Model excels on the training dataset but fails in real-world scenarios lacks practical usability.
- **Cross-Cultural and Language Variations:** Variations in expressing emotions across different languages and cultures pose an additional challenge to SER models.

Chapter 5

Standards Adopted

5.1 Design Standards

- Adhered to machine learning project design protocols
- Used UML-style flow diagrams for pipeline structure

5.2 Coding Standards

- Followed PEP8 standards for Python
- Modular code with reusable functions and clear comments

5.3 Testing Standards

- Used standard validation metrics
- Followed IEEE testing principles

Chapter 6

Conclusion and Future Scope

5.1 *Conclusion*

In this project, we developed a SER model using ML and DL techniques. The proposed models were trained and validated on the *IITKGP-SEHSC* dataset, utilizing features such as Mel Spectrogram, MFCC, ZCR and Chroma. Our results indicate that DL models outperform traditional ML classifiers. The highest accuracy achieved was **78.17%**, demonstrating the effectiveness of data-driven learning in SER.

While the system performed well, certain challenges persist. Emotional overlap, speaker variability, and cross-linguistic differences make emotion recognition a complex problem. The computational efficiency of transformer-based models remains a concern for the real-time applications. Nonetheless, findings suggest that advanced DL techniques can enhance the robustness and accuracy of SER models.

5.2 *Future Scope*

Future research will focus on the following areas to further improve SER systems:

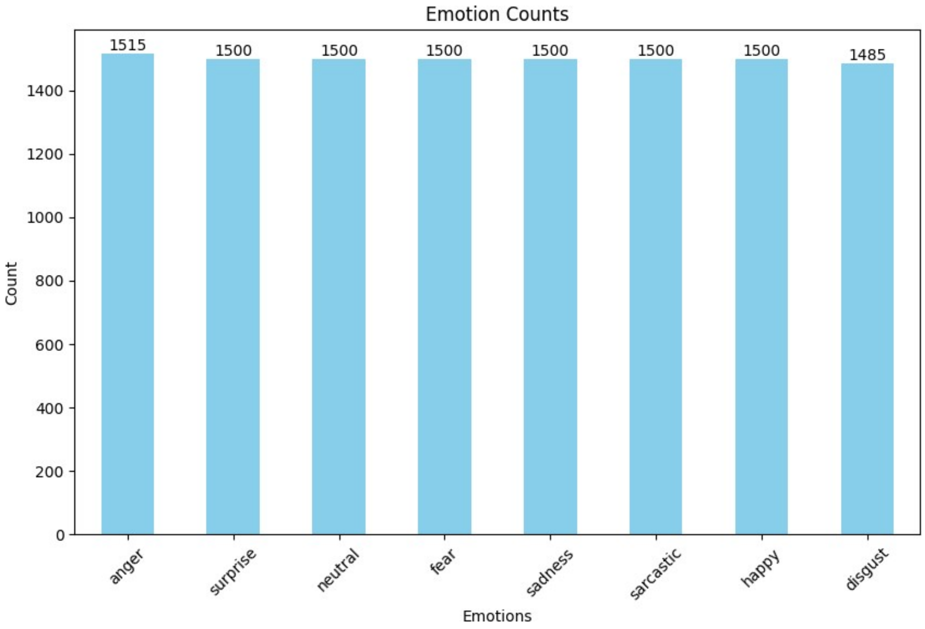
- **Dataset Expansion:** Increasing the diversity of datasets with multilingual and multi-accent speech samples to improve generalization.
- **Data Augmentation:** Using techniques such as time-stretch, pitch shift and noise injection to enhance model robustness.
- **Real-Time Implementation:** Optimizing models for deployment in real-time applications such as virtual assistants and mental health monitoring.
- **Multimodal Emotion Recognition:** Integrating textual and facial expression analysis alongside speech for a more comprehensive emotion recognition system.
- **Lightweight Models:** Developing efficient, low-computation models for mobile and embedded applications.
- **Cross-Cultural Adaptation:** Addressing variations in emotion expression across different languages and cultures to create globally adaptable models.

APPENDIX A

APPENDIX

A.1 *Dataset Description*

The dataset used for SER was *IITKGP-SEHSC*, which includes labeled Hindi speech samples categorized into emotions such as Anger, Disgust, Fear, Happy, Neutral, Sad, Sarcastic, and Surprise. Speech signal was sampled at 16 kHz.



References

[1] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "Iitkgp-sehsc : Hindi speech corpus for emotion analysis," in *2011 International Conference on Devices and Communications (ICDeCom)*, 2011, pp. 1–5.

[2] K. V. Krishna, N. Sainath, and A. M. Posonia, "Speech emotion recognition using machine learning," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 2022, pp. 1014–1018.

[3] S. Sharanyaa, T. J. Mercy, and S. V.G, "Emotion recognition using speech processing," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*, 2023, pp. 1–5.

[4] M. A. Gismelbari, I. I. Vixnin, G. M. Kovalev, and E. E. Gogolev, "Speech emotion recognition using deep learning," in *2024 XXVII International Conference on Soft Computing and Measurements (SCM)*, 2024, pp. 380–384.

SPEECH EMOTION RECOGNITION USING DEEP LEARNING

JAYA KISHAN PATEL
22053958

Abstract: Speech Emotion Recognition (SER) is a crucial domain in computer and human interaction, focused on improving machine comprehension of human emotions using audio signals.

. This project utilizes the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC dataset), which contains Hindi audio samples with 8 emotional variations.

Individual contribution and findings:

Research and selecting the dataset.
Loading and preprocessing the data.
Extract key audio features using Librosa(MFCC,Chroma,Mel spectrogram)
Visualizing sample audio

Individual contribution to project report preparation:

Basic introduction
Project planning

Individual contribution for project presentation and demonstration:

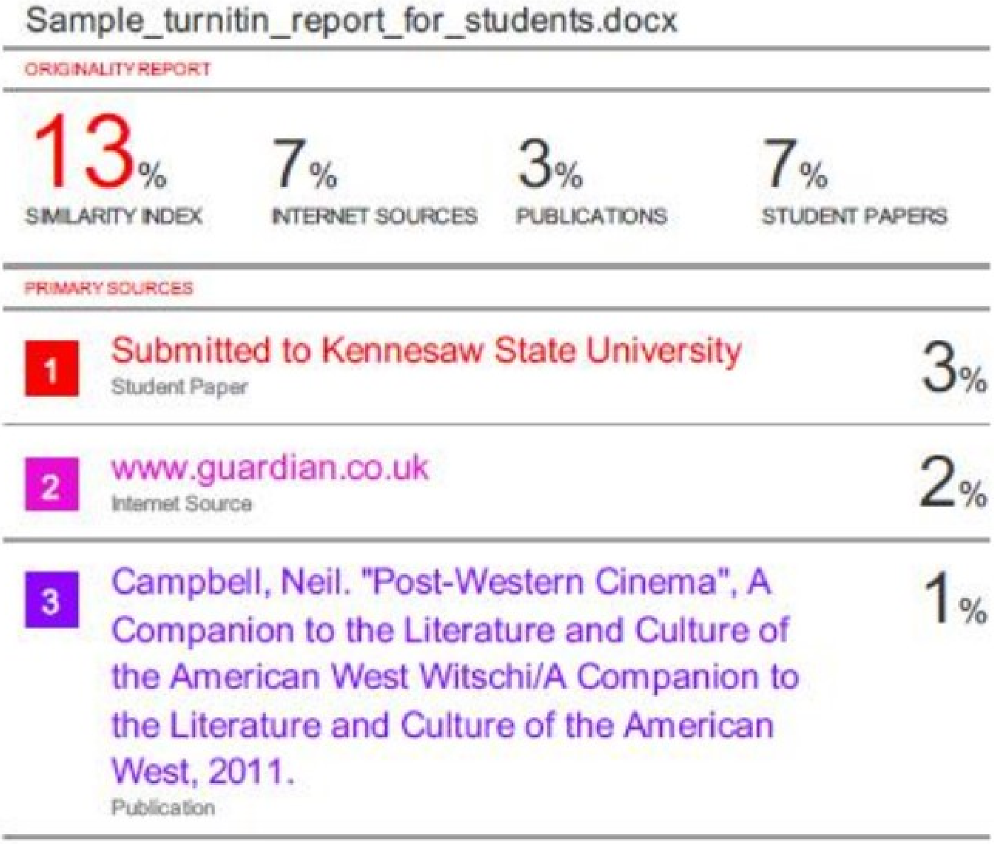
Introduction
Project planning
Project survey

Full Signature of Supervisor: ……………………………

Full signature of the student: ……………………………..

TURNITIN PLAGIARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)



Sample_turnitin_report_for_students.docx

ORIGINALITY REPORT

13%	7%	3%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Kennesaw State University Student Paper	3%
2	www.guardian.co.uk Internet Source	2%
3	Campbell, Neil. "Post-Western Cinema", A Companion to the Literature and Culture of the American West Witschi/A Companion to the Literature and Culture of the American West, 2011. Publication	1%