



Pokhara University

CITIZEN COLLEGE

PROJECT No.:2021-1-53-0376

VOICE EMOTION ANALYZER

BY

ARJUN THAPA

ROLL NO.: 21535062

A PROJECT REPORT

**SUBMITTED TO THE DEPARTMENT OF BACHELOR OF COMPUTER
APPLICATION**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR
THE DEGREE OF BACHELOR OF COMPUTER APPLICATION**

**DEPARTMENT OF BACHELOR OF COMPUTER APPLICATION
LALITPUR, NEPAL**

FEBRUARY,2026

Voice Emotion Analyzer

by

Arjun Thapa

Roll no.: 21535062

2021-1-53-0376

Project Supervisor

Er. Nishan Khanal

A project report submitted in partial fulfillment of the requirements for the degree of
Bachelor of Computer Application

Department of Bachelor of Computer Application

Pokhara University, Citizen College

Lalitpur, Nepal

February, 2026

COPYRIGHT ©

The author has agreed that the library, Department of Bachelor of Computer Application, Pokhara University, Citizen College, may make this project work freely available for inspection. Moreover the author has agreed that the permission for extensive copying of this project work for scholarly purpose may be granted by the professor(s), who supervised the project work recorded herein or, in their absence, by the Head of the Department, wherein this project work was done. It is understood that the recognition will be given to the author of this project work and to the Department of Bachelor of Computer Application, Pokhara University, Citizen College in any use of the material of this project work. Copying of publication or other use of this project work for financial gain without approval of the Department of Bachelor of Computer Application, Pokhara University, Citizen College and author's written permission is prohibited.

Request for permission to copy or to make any use of the material in this project in whole or part should be addressed to:

Head

Department of Bachelor of Computer Application

Pokhara University, Citizen College

Kumaripati, Lalitpur, Nepal

DECLARATION

I declare that the work hereby submitted for Bachelor of Computer Application at the Pokhara University, Citizen College entitled “**Voice Emotion Analyzer**” is my own work and has not been previously submitted by me at any university for any academic award. I authorize the Pokhara University, Citizen College to lend this project work to other institutions or individuals for the purpose of scholarly research.

Arjun Thapa

Roll no.: 21535062

2021-1-53-0376

February, 2026

RECOMMENDATION

The undersigned certify that they have read and recommend to the Department of Bachelor of Computer Application for acceptance, a project work entitled “**Voice Emotion Analyzer**”, submitted by **Arjun Thapa** in partial fulfillment of the requirement for the award of the degree of “**Bachelor of Computer Application**”.

Project Supervisor

Er. Nishan Khanal

Lecturer/Researcher

BCA Program Coordinator

Er. Nishan Khanal

Department of Bachelor of Computer Application, Citizen College

February, 2026

DEPARTMENTAL ACCEPTANCE

The project work entitled “**Voice Emotion Analyzer**”, submitted by **Arjun Thapa** in partial fulfillment of the requirement for the award of the degree of “**Bachelor of Computer Application**” has been accepted as a genuine record of work independently carried out by the student in the department.

Er. Nishan Khanal

Head of the Department

Department of Bachelor of Computer Application,
Citizen College.

February, 2026

LETTER OF APPROVAL

We certify that this project work (report) entitled “**Voice Emotion Analyzer**,” submitted by **Arjun Thapa**, has been examined and that we are satisfied with the project defense. In our opinion, the work is satisfactory in scope and qualifies as a project in partial fulfillment of the requirements for the degree of “**Bachelor of Computer Application**” under the Department of Bachelor of Computer Application, **Pokhara University**.

Project Supervisor

Er. Nishan Khanal
Lecturer/Researcher, Coordinator
Citizen College
February,2026

Project Board Member

Er. Dhiraj Bashyal
Lecturer/Researcher
Citizen College
February,2026

Examiner

February,2026

Principal

Hari Krishna Aryal,
Citizen College
February,2026

ACKNOWLEDGMENT

I would wish to thank the greatest of the great who contributed and supported this project and made it possible. I would like to show my utmost appreciation towards my supervisor, **Er. Nishan Khanal**, of **Citizen College**. Their precious advices, constructive criticisms, and support were essential in ensuring the completion of this work was successful. They influenced the course of my research and the quality of its results, based on their knowledge and positive advice. Moreover, they were the BCA Coordinator, and their thorough coordination and unambiguous decisive feedback made this entire project process much higher. My friends and fellow colleagues are also people I am grateful to. Their liberal advice, encouragement, and companionship during this time made the whole hard journey not only pleasant, but also rewarding in the end. And lastly, I would not be able to thank my family enough. It is my parents who, especially, inculcated in me unshaken faith and gave me a source of deep inspiration. It is because their consistent encouragement and constant emotional support became the invaluable base with the help of which I could attain my objectives.

Arjun Thapa

2021-1-53-0376

February, 2026

ABSTRACT

The voice enables one to understand human emotions, which is necessary in enhancing human intelligent interaction. In this project, a voice emotion analyzer is trained based on the convolutional neural network in order to differentiate between emotional states in response to audio. The voice characteristics such as the frequency patterns, energy, and variations in pitch are picked up and fed into the model to generate trusted input. The network is trained to identify happy, sad, angry, fear, disgust, neutral and surprise . Pitch analysis is also used in providing more context in gender of the speaker. According to the obtained results, the model achieves the mean accuracy of 83.89% The system is confident in its ability to predict emotions, which allows understanding the behavior and emotional state of the user better. The results of the experiments indicate that the given model qualifies as an accurate emotion recognition one in various situations. The sphere of application of this project in the field of mental health checks monitoring, interaction between humans and computers, and personalized services in order to be responsive to human emotions.

Keywords: Convolutional Neural Networks, Emotion Recognition, Frequency Patterns Pitch Analysis

TABLE OF CONTENTS

COPYRIGHT	iii
DECLARATION	iv
RECOMMENDATION	v
DEPARTMENTAL ACCEPTANCE	vi
LETTER OF APPROVAL	vii
ACKNOWLEDGMENT	viii
ABSTRACT	ix
TABLE OF CONTENTS	x
LIST OF FIGURES	xiv
LIST OF TABLES	xvi
LIST OF ABBREVIATIONS	xvii
1 INTRODUCTION	1
1.1 Background	3
1.2 Motivation	4
1.3 Problem Statement	5
1.4 Project Objectives	7
1.5 Scope of Project	7
1.6 Potential Project Applications	8
1.7 Feasibility Study	9
1.7.1 Technical Feasibility	9
1.7.2 Economic Feasibility	10
1.7.3 Operational Feasibility	10
1.7.4 Legal and Ethical Feasibility	10
1.7.5 Schedule Feasibility	11
1.7.6 Resource Feasibility	11
1.7.7 Market Feasibility	11
1.7.8 Social and Cultural Feasibility	12
1.8 Originality of the Project	13

1.9	Organization of Project	14
2	LITERATURE REVIEW	15
2.1	Literature Review I	15
2.2	Literature Review II	16
2.3	Literature Review III.....	17
2.4	Literature Review IV	18
2.5	Literature Review V	19
2.6	Literature Review VI	20
2.7	Literature Review VII.....	21
2.8	Literature Review VIII.....	22
2.9	Literature Review IX	23
2.10	Literature Review X	24
2.11	Summary of Literature Review	26
3	METHODOLOGY.....	29
3.1	Theoretical Formulations.....	29
3.2	Mathematical Modeling.....	30
3.2.1	Audio Feature Extraction	30
3.2.2	Convolutional Neural Network.....	30
3.2.3	Emotion Classification	31
3.2.4	Training Objective	31
3.3	System Block Diagram:	33
3.3.1	File Upload and Microphone Input	33
3.3.2	Audio Preprocessing	33
3.3.3	Feature Extraction.....	34
3.3.4	Deep Learning Model	34
3.3.5	Emotion Classification Output	35
3.3.6	Frontend Interface.....	35
3.4	Instrumentation Requirements	35
3.4.1	Explanation of Instrumentation Requirements	36
3.5	Dataset Explanation	37
3.5.1	Relevancy of the Dataset.....	37
3.5.2	Contents of the Dataset	37

3.5.3	Datasets Used	38
3.5.4	Dataset Representation	39
3.6	Description of Algorithm	42
3.6.1	Algorithm Pipeline	42
3.6.2	Stepwise Description	42
3.7	Elaboration of Working Principle	47
3.7.1	Voice Input	47
3.7.2	Pre-processing	47
3.7.3	Feature Extraction.....	47
3.7.4	Feature Dataset Formation	50
3.7.5	Model Building and Training	50
3.7.6	Model Evaluation	51
3.7.7	Prediction and Output	51
3.7.8	Testing with Live Voice	51
3.7.9	User Authentication with Firebase	51
3.7.10	Results and Interpretation.....	52
3.8	Verification and Validation Procedures	53
3.8.1	Verification	53
3.8.2	Validation.....	54
3.8.3	Classification Metrics Summary	56
4	RESULTS.....	57
4.1	Signup Page	57
4.2	Login Page	58
4.3	Landing Page.....	59
4.4	Activity Page	60
4.5	Live Audio Prediction	61
4.6	Emotion Distribution for Live Inputs.....	62
4.7	Live Audio Waveforms	62
4.8	Live Audio Spectrograms	63
4.9	Live Audio Suggestions	63
4.10	File Upload Prediction	64
4.11	File Upload Waveforms	64

4.12 File Upload Spectrograms	65
4.13 Emotion Distribution for Uploaded Files	65
4.14 File Upload Suggestions	66
4.15 History Page	66
4.16 Training and Validation Loss	67
4.17 Confusion Matrix	69
4.18 Best Case Scenario: Voice Emotion Analyzer	71
4.19 Worst Case Scenario: Voice Emotion Analyzer	72
5 DISCUSSION AND ANALYSIS	73
6 FUTURE ENCHANCEMENT	75
6.1 Unfulfilled Objectives and Remaining Challenges	75
6.2 Planned Improvements	75
7 CONCLUSION	77
APPENDIX A	
A.1 Project Schedule	78
A.2 Literature Review of Base Paper- I	79
A.3 Literature Review of Base Paper- II	80
A.4 Literature Review of Base Paper- III	81
A.5 Literature Review of Base Paper- IV	82
A.6 Literature Review of Base Paper- V	83
A.7 Literature Review of Base Paper- VI	84
A.8 Literature Review of Base Paper- VII	85
A.9 Literature Review of Base Paper- VIII	86
A.10 Literature Review of Base Paper- IX	87
A.11 Literature Review of Base Paper- X	88
REFERENCES	

LIST OF FIGURES

Figure 1.1	Voice Emotion Analyzer	4
Figure 3.1	Illustration of the CNN-based Voice Emotion Analyzer pipeline.	30
Figure 3.2	System Block Diagram of Voice Emotion Analyzer	33
Figure 3.3	Emotion distribution of all datasets.	40
Figure 3.4	Pipeline from audio input to output.	42
Figure 3.5	MFCC feature extraction workflow from raw audio to feature matrix.	44
Figure 3.6	CNN architecture for processing MFCC features and classifying emotions.....	45
Figure 3.7	Percentage of Accuracy from each Class	46
Figure 3.8	Mel spectrogram	48
Figure 3.9	MFCC.....	48
Figure 3.10	Chroma gram	49
Figure 3.11	Spectral Contrast.....	49
Figure 3.12	Zero Crossing Rate.....	50
Figure 3.13	Firebase Authentication	52
Figure 4.1	SignUp Page.....	57
Figure 4.2	Login Page	58
Figure 4.3	Landing Page	59
Figure 4.4	Activity Page	60
Figure 4.5	Live Audio Prediction	61
Figure 4.6	Emotion Distribution for Live Inputs	62
Figure 4.7	Live Audio Waveforms	62
Figure 4.8	Live Audio Spectrograms.....	63
Figure 4.9	Live Audio Suggestions	63
Figure 4.10	File Upload Prediction	64
Figure 4.11	File Upload Waveforms	64
Figure 4.12	File Upload Spectrograms	65
Figure 4.13	Emotion Distribution for Uploaded Files	65
Figure 4.14	File Upload Suggestions.....	66
Figure 4.15	History Page	66

Figure 4.16 Training and Validation Accuracy Curve	68
Figure 4.17 Training and Validation Loss Curve	68
Figure 4.18 Confusion Matrix	70
Figure A.1 Gantt Chart showing Project Timeline.	78

LIST OF TABLES

Table 2.1	Summary of Literature Review	28
Table 3.1	Instrumentation requirements	35
Table 3.2	Summary of all datasets used in the project.	39
Table 3.3	Gender distribution across the datasets.....	39
Table 3.4	Emotion-wise sample distribution.	40
Table 3.5	Partitioning of the dataset into training, validation, and test sets.	41
Table 3.6	Audio properties of the dataset.	41
Table 3.7	Classification Metrics for Voice Emotion Analyzer.....	56
Table 4.1	Best Case Scenario for Voice Emotion Analyzer	71
Table 4.2	Worst Case Scenario for Voice Emotion Analyzer.....	72

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
BCA	Bachelor of Computer Application
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CREMA-D	Crowd Sourced Emotional Multimodal Actors Dataset
DCT	Discrete Cosine Transform
DL	Deep Learning
DNN	Deep Neural Network
EMO-DB	Berlin Database of Emotional Speech
ERC	Emotion Recognition in Conversation
FPS	Frames Per Second
GB	Gigabyte
GPU	Graphics Processing Unit
HCI	Human-Computer Interaction
IEMOCAP	Interactive Emotional Dyadic Motion Capture
K-EmoCon	Korean Emotion Conversation Dataset
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LibROSA	Library for Audio and Music Analysis
LSTM	Long Short-Term Memory
MELD	Multimodal EmotionLines Dataset
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multilayer Perceptron
NLP	Natural Language Processing
NN	Neural Network
OS	Operating System
PRISMA-DTA	Preferred Reporting Items for Systematic Reviews and

	Meta-Analyses for Diagnostic Test Accuracy
RAM	Random Access Memory
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
RVM	Relevance Vector Machine
SAVEE	Surrey Audio-Visual Expressed Emotion Dataset
SEMAINE	Sustained Emotionally colored Machine human Interaction using Nonverbal Expression
SER	Speech Emotion Recognition
SSD	Solid State Drive
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TESS	Toronto Emotional Speech Set
UI	User Interface
USC	University of Southern California
UX	User Experience
VAD	Voice Activity Detection
WAV	Waveform Audio File Format

1 INTRODUCTION

Emotion recognition based on voice has raised more and more attention due to its relevance in human computer interaction. Smart Machines able to understand the emotional conditions of speech are able to react more wisely and thoughtfully, and therefore make communication more human and purposeful. The speech signals are, however, very complex and fluctuate. Different individuals, gender, style of speaking and recording have varying expressions of emotions. Additional features that make the correct identification of emotions even harder are background noises, microphone quality, and the peculiarities of the language. All of these considerations render voice emotion analysis one of the non-trivial problems, which involves sophisticated signal processing and machine learning algorithms. To solve this problem, the voice emotion analyzer project was created to come up with an automatic program which could analyze emotions based on speech records. The main task of the project was to create a system that would be able to hear an audio source and identify correctly the emotional state that the speaker would have. To attain this goal, there are a number of technical processes that were undertaken, such as dataset preparation, audio preprocessing, feature extraction, model selection, training, and the evaluation of performance. The stages were planned to make sure that emotional features that were found in speech were well learnt and remembered by the model. There are four popular emotional speech datasets that were used in training and testing. RAVDESS also includes the recordings of trained actors who demonstrated various emotions in a set of scripted sentences, and thus they gave high-quality and balanced emotional samples. TESS is a recording that allows hearing of female speakers who convey various emotions with the same level of pronunciation and clarity. SAVEE has emotional speech samples of men speakers, and therefore, it brings gender-specific variation to the dataset. Besides these publicly available datasets, a special dataset of about 250 voice samples was also created as a part of the project. The custom dataset was gathered through the invitation of the participants to record the speech samples describing various emotions. Both the spontaneous speech and the prepared sentences were employed to get the realistic emotional variation. Once the data had been collected, feature extraction was carried out to transform raw audio signals into numerical data which could be used in machine learning. The rate of change in sign of the audio signal was measured using ZCR to give information on the signal texture and

noisiness. The energy and intensity of speech that tends to change greatly depending on the emotional state was captured using RMS. MFCC were obtained to mimic human auditory perception in terms of spectral properties of speech representation. They were calculated with the help of the standard audio processing methods which are provided in the librosa library, and they were united in order to create a full set of features to classify emotions. Better generalization of the models was obtained by data augmentation on training samples. Noise injection was used to simulate real-world recording conditions and time stretching was used to adjust the speech rate without changing pitch and pitch shifting was used to adjust vocal pitch without changing duration. The techniques of augmentation gave rise to several variations of the same audio sample, a greater range of diversity in a dataset, and a lower chance of overfitting; after augmentation, the model was trained on a greater number of acoustic patterns. The classification of emotions was done based on CNN architecture because of its capability of extracting local patterns out of structured input data where by multiple convolution layers were trained to extract hierarchical features and normalization and pooling layers were used to stabilize the learning process and dimensional reduction. The option of dropout layers to prevent overfitting and dense layers to classify final emotions were added, and the loss function was categorical cross-entropy and Adam optimizer was applied to the training process using a constant number of epochs. Once the model had been trained, it was tested using validation data and performed with an accuracy of 83.89 percent, which suggested a consistent performance on seven categories of emotions; the emotion types like anger and happiness were identified more precisely since acoustic cues were more significant, but the fear and disgust were more confused since of the two, the fear had more similar speech patterns. Gender autodetection using pitch detection was also included as another option and proved to differentiate between male and female voice satisfactorily. To facilitate practical application, a web-based interface was created with the help of the Streamlit framework that enabled users to upload audio files or record the speech with a microphone, after which features were extracted, emotions were predicted, confidence scores were shown and visualization of both the waveform and spectrogram were offered with the emission of emotion-related suggestions offered to improve the interaction of the user, but still, creating and testing the emotion recognition model was the tipping point of the whole project.

1.1 Background

Emotions had formed a vital section of human communication which was not limited to verbal communication. All the utterances had described emotional information by altering tone, pitch, volume, and rhythm. Such vocal cues had helped human beings to comprehend feelings, intentions, and attitudes when communicating. Such emotional cues were the natural interpretations of human beings and they had been used to facilitate social interaction, empathy, and successful understanding among the individuals. The voice processing systems that had been traditionally developed were designed to translate spoken language to text. These systems had concentrated on the linguistic content and overlooked the emotional content that was contained in the voice signal. Consequently, the capacity of computers to decode the emotional context of voice had not been accomplished and this served as a strong differentiation between human communication skills and the machine-based interpretation. The artificial intelligence methods had developed to bring opportunities to narrow this gap. Scientists were now working on the ways of enabling the machines to identify and recognize emotions based on voice cues. The early emotion recognition systems had been based on manually derived acoustic features which included pitch, energy and frequency patterns. Even though these strategies had demonstrated early effectiveness, they had previously been unable to capture the variability and multi-dimensionality of emotional voice. In particular, the development of machine learning, and deep learning, had greatly enhanced the study of emotions. Deep learning models had already been able to learn meaningful patterns in audio data automatically rather than relying on predefined features completely. Convolutional neural networks were modified to identify voice representations in the form of spectrograms and proved to be able to detect differences in emotion better. Access to this field was also advanced by the availability of voice collections of emotion and the best audio processing tools. A stable baseline of emotion recognition system training and testing had been established on standardized data of emotion voice. Simultaneously, features extraction and signal processing had been made easier in the libraries of audio analysis. However, with these advancements, other issues like variability of speakers, environmental noise as well as cultural differences had not disappeared. Such challenges had pointed at the necessity of strong models and varieties of training data to acquire credible emotion recognition within the real-life setting.

1.2 Motivation

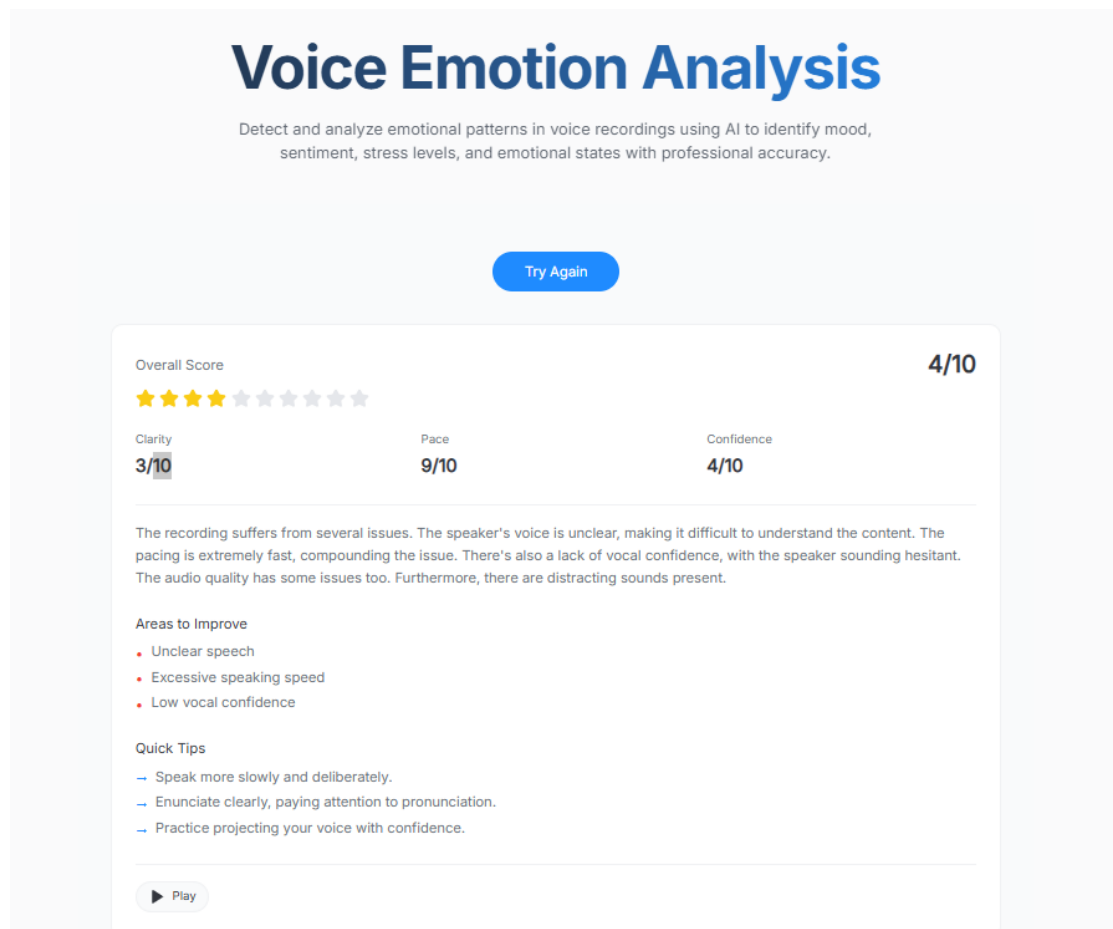


Figure 1.1: Voice Emotion Analyzer

The motivation of the Voice Emotion Analyzer project is to overcome limitations in existing human–computer interaction systems by enabling accurate emotion recognition through voice using artificial intelligence. The key motivating factors are outlined as follows:

- Lack of emotional awareness in current AI-based interaction systems.
- Challenges in accurately identifying human emotions from vocal signals.
- Limited naturalness and engagement in human–computer communication.
- Need to enhance interaction quality through voice-based emotion analysis.

1.3 Problem Statement

The development of an accurate and reliable voice emotion recognition system remains a significant challenge. Despite advancements in machine learning and artificial intelligence, existing systems face multiple limitations that hinder their accuracy, robustness, and applicability in real-world scenarios.

1. **Manual Feature Engineering Dependence:** Many systems rely on handcrafted features to detect emotions. This approach is time-consuming, less scalable, and often fails to capture subtle emotional cues in human voice.
2. **Preprocessing and Acoustic Pattern Extraction:** Identifying meaningful acoustic patterns from raw audio signals is challenging. Poor preprocessing can lead to noise sensitivity, reduced accuracy, and misclassification.
3. **CNN Architecture Design:** Designing an effective convolutional neural network for multi-class emotion classification is complex. Inadequate architecture may fail to distinguish between similar emotions.
4. **Dataset Limitations:** Emotional voice datasets are often small, imbalanced, or lack diversity. This limitation hampers model training, generalization, and fairness across different speakers and recording conditions.
5. **Computational Constraints:** Achieving reliable emotion recognition with low computational cost is difficult, especially for real-time applications.
6. **Spectral and Temporal Feature Analysis:** Extracting and leveraging spectral and temporal features effectively is essential for differentiating emotions with similar vocal patterns.
7. **Overfitting Risks:** Small and limited datasets increase the risk of overfitting, reducing the model's ability to generalize to new or unseen voice data.
8. **Speaker and Environment Variability:** Variations in speaker pitch, accent, or recording conditions can cause significant accuracy fluctuations, challenging the robustness of the system.

9. **Model Validation Challenges:** The absence of large, publicly available emotion datasets complicates model validation and benchmarking.
10. **Real-World Generalization:** Extending emotion recognition models to real-world conversational voice is challenging due to background noise, spontaneous speech, and emotional subtlety.

To address the above challenges and guide the development of an effective Voice Emotion Analyzer, the following research questions are formulated:

1. How can emotions be identified from voice without manual feature engineering?
2. Which preprocessing methods best extract meaningful acoustic patterns from raw audio?
3. How can an effective CNN architecture be designed for multi-class emotion classification?
4. What strategies can mitigate dataset limitations and imbalance in emotional voice recordings?
5. How can reliable emotion recognition be achieved with low computational cost?
6. What role do spectral and temporal features play in separating similar emotions?
7. How can overfitting on small datasets be prevented?
8. How can accuracy variation across speakers and recording conditions be minimized?
9. How can models be validated effectively without large public datasets?
10. What strategies can generalize emotion recognition to real-world conversational voice?

1.4 Project Objectives

1. Major

- To implement a machine learning model that accurately detect human emotions from voice.

2. Minor

- To analyze and extract useful features from audio files using Python libraries for emotion classification.
- To test and evaluate deep learning model to achieve the accuracy in emotion recognition.

1.5 Scope of Project

The project of the Voice Emotion Analyzer creates a system that can be used to recognize emotions based on voice signals and offer the reliable means of interpreting the emotions in human beings. Audio input falls into seven different emotions including angry, disgust, fear, happy, neutral, sad and surprise. The basis of analysis had been provided by the extraction of audio features such as zero crossing rate, energy and at Mel-frequency cepstral coefficients. The noise addition, pitch variation, and time stretching data augmentation techniques had been used to enhance the model and enhance its capability in generalizing to various voice patterns and speakers. The convolutional neural network had been trained on four publicly available emotional voice datasets as well as 250 recorded voice samples. The trained model runs on a web interface, which takes uploaded audio files or real time recording. The interface shows determined emotions with confidence scores, and visualizes audio of a speaker and spectrograms and audio waveforms, and, above all, the gender of the speaker. Preprocessing, such as noise reduction, signal normalization, and framing, had been made to improve the clarity of features. Spectral and temporal characteristics enable the model to discriminate similar emotions and CNN records both local and global audio patterns, so no misclassifications can occur. The interface gives instant feedback and delivers the results in a friendly and comprehensive way.

Although the system has its capabilities, it has some shortcomings. The processing of audio input is only performed, and facial expressions or text-based recognition

of emotions are not considered. The audio streams are not continuous. With few Nepali samples and mostly with English recording, the training data lessen the accuracy on other languages. Only recognized emotions are the seven trained ones and the person speakers are not distinguished. The system is also less accurate when the audio is extremely noisy or with spontaneous speech rather than acted samples. Complex emotional expression, sarcasm and cultural variations are not perceived either. Privacy and consent protection was not implemented, and the sensitive user information is not handled. The difference in the pitch of the speaker, tone, accent and recording environment causes fluctuations in performance. The limitations underscore the need to be cautious with data preparation, extracting features, and evaluating them in order to achieve a dependable and understandable voice emotion recognition system.

1.6 Potential Project Applications

1. **Mental Health Monitoring:** Voice-based emotion recognition assists mental health professionals by providing objective measurements of patient emotions over time. Regular recordings identify mood patterns, detect indicators of depression or anxiety, and monitor treatment effectiveness. The system alerts healthcare providers to concerning emotional changes while reducing reliance on subjective self-reporting.
2. **Customer Service Analysis:** Call centers utilize emotion detection to assess caller sentiment in real-time. Distressed callers are highlighted for attention, and supervisors obtain emotional analytics for quality assurance. Post-call analysis reveals service pain points and informs staff training based on emotional interaction patterns.
3. **Educational Technology:** Adaptive learning platforms gauge student engagement and frustration during online instruction. Detection of confusion or disengagement allows dynamic adjustment of content difficulty and presentation style. Educators gain feedback on emotional responses to teaching methods, facilitating early support for struggling students.
4. **Human-Computer Interaction:** Voice-controlled interfaces, virtual assistants, and interactive systems benefit from understanding user emotions. The system

adapts responses based on detected feelings, enabling empathetic interactions and improving overall user experience.

5. **Entertainment and Media Analysis:** Emotion recognition is applied to analyze audience reactions to audio content such as podcasts, audiobooks, and voice performances. This helps content creators understand engagement levels, emotional impact, and listener preferences.
6. **Research in Speech and Linguistics:** The system supports academic studies by analyzing emotional expression in speech. Researchers can study voice patterns, cross-cultural differences, and the effect of intonation and pitch on emotion, contributing to studies in linguistics, psychology, and AI.
7. **Voice-Based Accessibility Tools:** Emotion detection enhances accessibility for users with communication difficulties. Applications include assisting speech therapy, supporting individuals with autism spectrum disorder, or providing feedback for users practicing emotional expression in speech.

1.7 Feasibility Study

The development of the Voice Emotion Analyzer, an automated system for emotion recognition from human voice, requires a comprehensive feasibility analysis. This section evaluates the project from multiple perspectives to ensure technical viability, economic sustainability, operational efficiency, legal compliance, and timely completion.

1.7.1 Technical Feasibility

The system relies on modern AI and deep learning technologies, open-source tools, and well-supported libraries to ensure technical feasibility.

- **Audio Processing and Feature Extraction:** Python libraries such as LibROSA and NumPy handle audio preprocessing, feature extraction, and analysis efficiently.
- **CNN-Based Classification:** TensorFlow and PyTorch provide complete support for building and training the convolutional neural network.
- **Web Interface Integration:** The system operates through a web interface, accepting uploaded audio files and live recordings, displaying emotions, confidence

scores, audio waveforms, spectrograms, and speaker gender.

- **Hardware Requirements:** The system runs smoothly on a standard laptop with 8GB RAM and SSD storage, and audio signals are lightweight to process.

1.7.2 Economic Feasibility

The project is economically feasible because it utilizes freely available datasets and open-source libraries.

- **Cost Efficiency:** RAVDESS, SAVEE, and other public datasets are used for training and testing, eliminating the need for paid resources.
- **Minimal Hardware Expense:** Existing laptops and standard computing resources were sufficient for development and operation.
- **Sustainable Development:** All software tools are open-source, resulting in negligible financial burden and suitability for academic and research purposes.

1.7.3 Operational Feasibility

The system demonstrates operational feasibility due to its straightforward workflow and user interaction.

- **Workflow Simplicity:** Users provide a voice sample, and the system automatically handles preprocessing, feature extraction (MFCC, Chroma, Spectral Contrast), and CNN-based emotion prediction.
- **User-Friendly Interface:** Emotion results, confidence scores, and visualizations are presented clearly, ensuring interpretability.
- **Practical Applications:** The system supports emotional monitoring, customer service analysis, and voice-based assistance tools effectively.

1.7.4 Legal and Ethical Feasibility

The project maintains legal and ethical compliance throughout its development.

- **Data Privacy:** The system uses datasets allowed for academic research and does not store personal data unless explicitly required.
- **Licensing Compliance:** No copyrighted materials are used, ensuring legal safety.
- **Ethical Considerations:** The system respects privacy standards and avoids unauthorized use of sensitive information.

1.7.5 Schedule Feasibility

The project timeline aligns with academic and research deadlines.

- **Development Phases:** Dataset preparation, feature extraction, model training, interface creation, and testing have been completed according to a structured workflow.
- **Progress Monitoring:** Each phase had defined milestones, ensuring smooth execution and timely project completion.

1.7.6 Resource Feasibility

The system's resource requirements are manageable.

- **Human Resources:** Development, model training, and interface creation were completed by a small team of developers and testers.
- **Technical Resources:** Existing computing devices and open-source tools were sufficient for all project tasks.

1.7.7 Market Feasibility

There is a growing demand for intelligent, emotionally aware systems.

- **Target Users:** AI researchers, developers, and organizations interested in human-computer interaction, emotional monitoring, and accessibility solutions.
- **Competitive Advantage:** The system provides multi-emotion detection, gender identification, and visual feedback, differentiating it from simple audio analysis tools.

1.7.8 Social and Cultural Feasibility

The system aligns with modern technological and social practices.

- **Emotional Awareness:** Supports better interaction between humans and machines, enhancing engagement and understanding.
- **Cultural Adaptability:** While trained mostly on English and some Nepali samples, the system demonstrates flexibility and potential for expansion to other languages with similar voice patterns.

1.8 Originality of the Project

The Voice Emotion Analyzer project is unique as it develops a smart voice-based emotion recognition system that can interpret human emotions in real time. In contrast to traditional systems that rely on limited datasets or manual feature extraction, this project focuses on fairness, accuracy, and reliability of emotion recognition across different speakers and recording conditions.

The primary contribution of this project lies in the development of a system that **combines four publicly available emotional voice datasets with a self-collected corpus of 250 labeled samples**, creating an extensive and diverse training set that improves the **generalization capability of the model**. Furthermore, the system applies **data augmentation techniques such as noise injection, time stretching, and pitch shifting** to reproduce real-world recording variations and enhance robustness. A **custom convolutional neural network architecture** utilizes **acoustic features instead of raw audio** to classify **seven emotions angry disgust fear happy neutral sad and surprise**. Another notable contribution is the **integration of pitch-based gender detection with emotion recognition**, enabling simultaneous analysis of the speaker's gender and emotional state.

1.9 Organization of Project

This project report has been structured into seven main chapters, followed by the appendix and references. The first chapter introduces the research topic, including the background, motivation, problem statement, objectives, and scope of the study. Chapter 2 provides a detailed review of the literature, examining important publications and identifying gaps that this research addresses. Chapter 3 explains the methodology, covering the system design, algorithms, data processing methods, and technologies used in the development. Chapter 4 presents the results of the system, including performance measurements, visual outputs, and key observations. Chapter 5 discusses and analyzes the results, comparing them with theoretical expectations and existing systems, while highlighting limitations. Chapter 6 outlines potential future enhancements, describing improvements and additional functionalities that could be implemented. Chapter 7 concludes the report by summarizing the major findings, assessing the overall effectiveness of the system, and reflecting on its contributions. The appendix contains the project timeline and detailed literature reviews of base papers, followed by a complete list of references to all cited research works.

2 LITERATURE REVIEW

2.1 Literature Review I

The paper titled Voice Emotion Recognition in Conversations Using Artificial Intelligence, A Systematic Review and Meta Analysis by Leontios Hadjileontiadis was a comprehensive and well-organized analysis of the development of research in the field of conversational voice emotion recognition[1]. This review was conducted in contrast to traditional studies that were based on isolated utterances, whereas natural human conversation involved emotional expression that depended on context, interaction patterns, and speaker relationships[1]. The research emphasized that emotion recognition in real conversations was more complex than emotion detection from single sentences because emotions evolved over time, reacted to other speakers, and were highly situation-dependent[1]. A large body of research literature was systematically structured and evaluated through a rigorous review methodology[1]. The review provided a broad overview of artificial intelligence methods used in conversational emotion recognition, including classical feature-based approaches, temporal modeling using recurrent neural networks, spectral representations with convolutional neural networks, long-range dependency modeling through attention mechanisms, and contextual sequence learning using transformer-based architectures. The meta-analysis evaluated performance improvements across these approaches and demonstrated that deep learning models significantly outperformed traditional methods[1]. The development of multimodal conversational datasets was also highlighted in the review[1]. These datasets combined audio, textual, visual gestures, and physiological signals, enabling more accurate emotion recognition[1]. Multimodal systems achieved superior performance because conversational emotions were often influenced by subtle contextual cues that could not be captured through voice alone. However, several challenges were identified, including the scarcity of high-quality conversational datasets, inconsistencies in emotion annotation schemes, cross-cultural variability in emotional expression, and difficulties in modeling overlapping emotions during conversations[1]. The review critically discussed the limitations of existing systems, noting that models trained on acted datasets with exaggerated emotions performed poorly in real conversational environments characterized by background noise, informal speech, interruptions, sarcasm, and uncertainty. Context modeling, speaker diarization, and sequence-based learning were identified as essential components for

improving system robustness. Future research directions included cross-cultural adaptation, context-aware emotion tracking, ethical considerations in emotion monitoring, unified benchmarking standards, and real-time adaptive systems, with applications in virtual assistants, call centers, and mental health monitoring contributing to more natural human–computer interaction[1].

2.2 Literature Review II

The research paper entitled Emotion Recognition on Call Center Voice Data explored the significance of voice emotion recognition in improving customer service activities in call centers[2]. The analysis was based on real customer interactions rather than controlled or acted speech, making the study more practical in real-world applications. Call center conversations were often characterized by emotional strain, dissatisfaction, indecisiveness, and urgency, which made emotion detection an essential tool for understanding customer needs[2]. As demonstrated in the study, identifying customer emotions during interactions enabled agents to respond more empathetically and effectively. Unlike earlier emotion recognition studies that relied on clean datasets, this research addressed challenges associated with noisy and spontaneous speech[2]. Authentic call recordings from Turkish mobile operators were used, incorporating background noise, varied speaking styles, and interrupted speech, thereby closely simulating real call center environments and increasing the difficulty of emotion recognition. This realistic analysis allowed the study to present a practical evaluation of system performance and usability[2]. The system employed deep learning techniques to classify emotional states into positive, negative, and neutral categories[2]. This simplified classification scheme was suitable for operational environments where fast and direct emotional feedback was required. An accuracy of 0.91 was reported, demonstrating that deep learning models could successfully learn emotional patterns even from complex and unstructured voice data[2]. The system provided emotional intelligence support to call center employees, helping them adapt their communication strategies during customer interactions. The authors concluded that integrating emotion recognition into call center workflows improved customer satisfaction and service quality[2]. The study also emphasized the importance of language-specific research, as the system was developed for Turkish speech, a language that is underrepresented in emotion recognition studies. The findings suggested that future systems could be extended to multilingual environments and enhanced with finer-

grained emotion categories for more detailed emotional analysis[2]. Another significant contribution of the research was the investigation of feature selection and preprocessing techniques to optimize emotion classification under real-world conditions[2]. Noise filtering, voice segmentation, and spectral feature extraction methods were applied to improve input data quality. The study also identified speaker variability as a factor influencing model performance and discussed approaches to reduce bias across different voice types and accents[2]. These techniques supported the development of more robust and generalizable emotion recognition systems that could be effectively deployed in real call center application settings.

2.3 Literature Review III

The paper titled Emotion Detection via Voice and Voice Recognition, written by Rohit Rastogi, Tushar Anand, Shubham Sharma, and Sarthak Panwar, was a thorough investigation of detecting human emotions through the spoken voice cues[3]. The paper has started by highlighting the increasingly critical role of emotion-conscious computing in the context whereby intelligent systems were interacting with human beings more frequently. Voice being natural and impulsive had substantial acoustic signals that disclosed the emotional condition of a person even when the verbal message itself was neutral[3]. The objective of the study was to develop a strong model that would explain these emotional cues through the combination of state-of-the-art signal processing and machine learning methods. The study has described the audio preprocessing steps to be taken in recognizing emotions in detail[3]. Voice cues that were observed across different people contained huge differences in tone, volume, noise in the background, and the manner of speaking. Direct processing of audio content in its pure form resulted in the model paying attention to unrelated anomalies and not emotional content[3]. In order to solve this, the pipeline was utilized that incorporated noise removal, silence clipping, feature scaling, and transformation of audio signals into homogeneous segments of length. The result of this process was pure and reliable representations of voice that had a direct effect on the accuracy of classifiers[3]. The paper has reinforced the role of feature extraction[3]. There was a lot of use of Mel Frequency Cepstral Coefficients and other spectral and temporal descriptors. These features measured vital characteristics of emotions including change in energy, change in pitch dynamics, harmonic structure, and frequency distribution[3]. Emotions like anger were linked with more energy, more

harmonics, and broader frequencies, whereas sadness was reflected in less energetic, less harmonic, and less high-pitched voice. The emotion recognition model was based on these differences in acoustics[3]. Classical machine learning algorithms as well as deep learning architectures were tried[3]. Conventional algorithms such as Support Vector Machines and Random Forest classifiers performed well using well-designed features. Nevertheless, deep learning models, especially convolutional neural networks, had a much higher performance as compared to the classical models[3]. The space patterns that were found to be related to emotional tone in convolutional layers were automatically identified and the use of manual feature engineering was minimized. The findings showed that certain emotions were more classifiable as compared to others[3]. Emotions with extreme acoustic profiles such as anger and happiness were better recorded, whereas similar acoustic emotions such as calm and sad caused confusion[3]. The analysis indicated the need of bigger more varied datasets in order to enhance generalization. Human-computer interaction, personalized digital assistants, driver monitoring systems, and emotion-aware recommendation engines were also used in practice[3]. The work laid a solid foundation to contemporary voice-based affective computing.

2.4 Literature Review IV

The paper Evaluated the Impact of Voice Activity Detection on Voice Emotion Recognition in Autistic Children aimed at enhancing the rate of emotion recognition during child voice especially with autism[4]. The voice of autistic children also did not correspond to usual patterns of speech and contained uneven timing, abnormal intonation, and prolonged pauses. These attributes complicated emotion recognition and decreased the efficiency of regular emotion recognition systems[4]. The authors focused on preprocessing methods, especially voice activity detection, to extract useful voice segments. Silence, background noise, and unwanted audio contents were eliminated with voice activity detection and then emotion analysis occurred[4]. Such a step was necessary for child voice, in which emotional indications were subtle and easily concealed by noise or non-vowel sounds. The study showed through experimental assessment that the reliability of emotional feature extraction with voice activity detection had been enhanced[4]. More stable results of emotion recognition were achieved through better segmentation of voice. The study also revealed that preprocessing was vital in the overall performance of voice emotion recognition systems, particularly in dealing with

sensitive and varying voice data[4]. The paper emphasized the need for special emotion recognition systems responsive to vulnerable groups like autistic children. Ethical issues regarding data gathering, system implementation, and interpretation of emotional cues were also highlighted[4]. Another contribution made by the study was demonstrating the effect of adaptive preprocessing on system robustness[4]. The results indicated that voice activity detection, when carefully tuned, improved the extraction of subtle emotional features even with difficult recordings. The paper also provided guidance on designing child-centered emotion recognition systems and suggested that more customized approaches would enhance reliability and usability in therapeutic, educational, and clinical contexts[4]. These findings highlighted the significance of preprocessing in the creation of useful, specialized emotion-aware technologies that can be applied to sensitive groups.

2.5 Literature Review V

The system introduced by Sadil Chamishka, Ishara Madhavi, Rashmika Nawaratne, Daminda Alahakoon, Daswin De Silva, Naveen Chilamkurti, and Vishaka Nanayakkara was a voice-based emotion recognition system which worked in real time to detect temporal variations in emotion in voice using recurrent neural networks and refined feature modelling[5]. The procedure was based on the creation of RNN constructions that preserved the sequential dependence of acoustic characteristics, enabling the system to acquire dynamic variation of vocal expressions much more effectively than inherent classifiers. The architecture operated over voice streams and had been scaled to low-latency real-time applications, which was highly useful in interactive systems[5]. To increase emotional discrimination, feature modelling was applied to better represent prosodic and spectral voice features that were inputted into the RNN classifier[5]. This allowed detection of transitioning emotional states with greater precision by combining temporal modelling with advanced feature extraction because of continuous voice input. The technique was tested on benchmark emotional voice datasets, and the experimental findings revealed significant improvement in recognition accuracy compared to machine learning models trained in baseline conditions[5]. The RNN-based structure was more accurate than traditional classifiers and demonstrated the advantages of temporal dependency modelling. The system was qualitatively stable in processing continuous voice streams and tracking changes in emotions in real time, which was necessary for practical human-computer interaction use[5]. Some of the strengths identified by the study

included the design of a real-time architecture, effective temporal modelling through recurrent neural networks, and enhanced feature representation resulting in superior recognition of emotions[5]. The low-latency nature of the framework made it highly practical, and benchmark validation empirically supported the research methodology. Limitations included limited multilingual coverage, susceptibility to real-world noise, lack of multimodal inputs such as facial expressions or physiological signals, and absence of robustness analysis in adverse acoustic settings, which constrained large-scale use[5]. Another contribution of this study was demonstrating the benefits of ageing temporal modelling with stronger feature representation in continuous emotion recognition[5]. The research provided guidance on designing real-time interactive voice emotion systems and highlighted future directions, such as covering more languages, integrating multimodal signals, and testing robustness in unfavorable acoustic environments. The results facilitated the development of robust, precise, and practically implementable emotion-sensitive systems[5].

2.6 Literature Review VI

The study titled A Cross Cultural Investigation of Emotion Inferences of Voice and Voice: Implications of Voice Technology was conducted by Klaus R. Scherer, in which the perception of vocal emotions across different cultural backgrounds had been systematically examined[6]. The expression of emotions through voice had been considered a complex interaction of physiological mechanisms, learned cultural conventions, linguistic habits, and individual expressive styles. Through this investigation, an attempt had been made to determine whether emotional cues conveyed through voice had been universally interpreted or whether they had been significantly influenced by cultural conditioning [6]. Voice samples representing a range of emotional states had been collected and presented to listeners belonging to multiple cultural groups[6]. Participants had been instructed to identify emotions based solely on acoustic cues such as pitch variation, intensity, speech rate, rhythm, and intonation patterns, without access to contextual or semantic information[6]. This approach had ensured that emotional inference was derived purely from vocal characteristics rather than external factors. The results had indicated that certain emotional patterns had been consistently recognized across cultures. High pitch, rapid tempo, and increased intensity had been commonly associated with emotions such as fear and excitement, while low pitch, reduced energy, and slower tempo had

been largely linked with sadness[6]. These findings had supported the hypothesis that some vocal expressions of emotion had biological foundations shared across human populations[6]. The notable cultural variations had also been identified[6]. Emotional interpretations had differed depending on social norms, linguistic traditions, and cultural exposure[6]. For instance, louder vocal expressions had been perceived as anger in some cultures, whereas the same acoustic patterns had been interpreted as confidence or enthusiasm in others[6]. These discrepancies had demonstrated that emotional perception through voice had not been entirely universal[6]. The implications for voice-based technologies had been strongly emphasized in the study[6]. It had been suggested that emotion recognition systems trained on culturally limited datasets were likely to perform poorly when deployed across diverse populations[6]. The research had highlighted the necessity of incorporating culturally diverse training data and adaptive mechanisms capable of adjusting emotional interpretations based on user background[6]. Another significant contribution of the study had been the emphasis on sociocultural awareness in emotion recognition system design[6]. The findings had shown that accurate emotional interpretation could not rely solely on acoustic features without consideration of cultural context. By integrating cultural sensitivity into algorithmic models, both recognition accuracy and user acceptance had been improved. This work had provided foundational insights for the development of globally applicable and socially aware voice emotion recognition systems[6].

2.7 Literature Review VII

The paper titled Multimodal Integration of Emotional Signals of Voice, Body and Context had examined how emotional understanding had been influenced when voice, body movement, and contextual information had been jointly considered in communication systems[7]. It had been argued that emotional expression had rarely been conveyed through voice alone and had instead been formed through an integrated combination of vocal tone, physical gestures, and situational cues[7]. The study had been positioned within the domain of human-robot interaction, where emotional intelligence had been considered essential for social acceptance and effective interaction[7]. Emotional perception accuracy had been evaluated under conditions where emotional cues had been either congruent or incongruent across modalities[7]. It had been observed that recognition performance had been significantly improved when voice, body language, and contex-

tual signals had conveyed consistent emotional information, whereas mismatched cues had caused ambiguity and reduced trust in robotic systems[7]. User attitudes toward robots had been analyzed, and robots displaying multimodal emotional consistency had been perceived as more intelligent, reliable, and socially engaging[7]. The findings had indicated that emotionally coherent multimodal behavior had positively influenced user comfort, interaction quality, and long-term acceptance[7]. Practical deployment considerations had also been addressed, where multimodal emotion systems had been shown to perform more robustly under dynamic and unpredictable real-world conditions[7]. Additionally, it had been suggested that future systems should incorporate extended sensory channels and long-term user interaction studies to enhance emotional adaptability and realism[7]. Overall, the study had reinforced that multimodal emotion recognition had better reflected natural human communication patterns and had been critical for developing socially intelligent artificial agents[7].

2.8 Literature Review VIII

The study titled Voice Based Emotion Recognition with Convolutional Neural Networks in Companion Robots had focused on integrating voice-based emotion recognition capabilities into companion robotic systems designed for close human interaction[8]. Companion robots had been widely applied in healthcare assistance, elderly support, child interaction, and personal companionship, where emotional awareness had been essential for natural engagement[8]. The research objective had been centered on constructing a convolutional neural network architecture capable of classifying emotional states using voice signals[8]. Voice recordings had been transformed into spectrogram representations, which had been treated as two-dimensional inputs enabling spatial and temporal feature learning[8]. Emotional characteristics such as high energy and spectral sharpness for anger and reduced frequency variation for sadness had been implicitly learned by convolutional filters without manual feature extraction[8]. Data preprocessing techniques such as normalization, segmentation, and augmentation had been applied to improve generalization across speakers and recording environments[8]. The CNN model had been composed of stacked convolutional and pooling layers followed by fully connected layers responsible for emotion classification[8]. Training had been performed using backpropagation on labeled emotional voice samples, and improved robustness against noise had been demonstrated when compared to traditional machine

learning approaches[8]. Real-time deployment within robotic platforms had enabled adaptive behavioral responses such as tone adjustment and interaction modulation based on detected emotions[8]. It had also been emphasized that performance had been further enhanced when diverse accents, background noise, and spontaneous speech patterns had been included during training[8]. The study had concluded that combining voice emotion recognition with additional modalities could further strengthen emotional intelligence in companion robots operating in complex social environments[8].

2.9 Literature Review IX

The paper by Fatemeh Noroozi, Tomasz Sapinski, Dorota Kaminska, and Gholamreza Anbarjafari introduced a vocal-based emotion recognition framework in which ensemble learning was implemented to model paralinguistic voice features and classify emotions accurately[9]. The basic algorithm used prosodic and spectral data, including pitch, intensity, formants, and other paralinguistic representations as input to a random forest classifier composed of multiple decision trees. This ensemble architecture was designed to provide greater robustness and generalization by aggregating predictions from multiple weak learners rather than relying on a single classifier[9]. The system was tested on the SAVEE emotional voice database using two validation strategies: leave-one-out cross-validation and 9-fold cross-validation[9]. Quantitative results reported an average score of 76.28 across six emotion categories. Happiness achieved the highest recognition rate of 78, indicating that positive affective states were more easily recognized. The effectiveness of the ensemble learning approach was demonstrated through its superior performance over Linear Discriminant Analysis and deep neural network baselines on the same dataset[9]. Qualitative analysis revealed that the system remained stable in classifying emotional categories and that strength improved when multiple decision trees were aggregated. The framework was therefore considered generalizable to multi-class emotion categorization and suitable for emotion-sensitive human-computer interaction applications[9]. Strengths of the study included a well-structured methodological process based on ensemble learning that enhanced classification stability and minimized overfitting[9]. Emotional separability improved through the use of discriminative paralinguistic features, and reliability increased with cross-validation strategies. Empirical evidence of the approach's superiority was presented by comparison with LDA and DNN baselines. The model demonstrated practical applicability in emotion-sensitive

HCI systems[9]. Limitations were also observed[9]. Experiments were conducted on a single small benchmark dataset, limiting generalizability. Reliance on acted emotional voices reduced ecological validity. Multimodal emotion cues such as facial expressions or physiological signals were not integrated, restricting the scope of emotion representation. Certain emotional categories, such as fear and surprise, had low recognition accuracy, and real-time deployment was not tested, leaving system latency and practicality unexamined[9]. Another point raised by the research was the potential of combining this ensemble learning structure with dynamic adaptive systems[9]. The emotion recognition system could be enhanced by incorporating real-time user responses and lifelong learning to better handle different speaker traits, spontaneous facial expressions, and contextual variations. This adaptation could improve reliability, practicality, and applicability in various real-life human-computer interaction scenarios[9].

2.10 Literature Review X

The article titled Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction by Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah focused on the important issue of variability in speaker voices in emotion recognition systems[10]. The authors stated that voice generated by different people was inherently different because of physiological variations in vocal tract length, habitual speaking styles, accent, gender, age, and emotional expressiveness. These differences disrupted the model's ability to focus solely on emotional cues[10]. When a system was trained on data from a small group of speakers, it tended to overfit to particular speaker characteristics instead of learning emotion-related features. Speaker variability was identified as a key reason why continuous emotion prediction remained unstable in cross-speaker conditions[10]. The study proposed a speaker normalization method based on factor analysis[10]. The technique decomposed voice elements mathematically into components representing the speaker and components conveying emotion. This separation allowed the model to focus on variations in arousal, valence, and dominance, the three continuous dimensions typically used to describe emotional states. Unlike categorical emotion classification, continuous emotion prediction monitored the strength of emotions over time, which was claimed to be more suitable for real-world applications such as dialog systems, behavioral analytics, and affective monitoring[10]. Voice features were decomposed into latent variables using factor analysis[10]. This enabled speaker-specific characteristics to be distinguished and

their influence minimized. Emotional cues became more evident and consistent across different speakers. Large-scale experiments on benchmark datasets with diverse speakers and emotional levels demonstrated that speaker normalization significantly improved the smoothness and accuracy of predicted emotional trajectories[10]. A notable strength of the study was the discussion of how subtle shifts in pitch or energy due to speaker identity could be confused with emotional changes[10]. Such variations were more problematic for continuous emotion models compared to categorical models, making normalization particularly important. Unnormalized models produced jagged, unpredictable, and noisy emotional predictions, whereas normalized models generated more stable and interpretable emotional curves[10]. The study also highlighted that speaker normalization frameworks could be integrated with deep learning architectures to further improve cross-speaker adaptability[10]. Neural network pipelines that included normalization allowed emotional representations to be less dependent on individual speaker characteristics, leading to more effective deployment in multiethnic populations. This design emphasized that preprocessing and normalization were as crucial as model architecture for achieving high-quality continuous emotion recognition[10].

2.11 Summary of Literature Review

Author/Year	Focus	Methodology	Key Findings	Contributions
Hadjileontiadis (2025)	voice emotion recognition in conversations	Systematic review and meta-analysis of CNN, RNN, Transformer, and attention-based models	Deep learning models significantly outperform traditional approaches; contextual and multimodal information improves emotion understanding	Provides a comprehensive roadmap for conversational emotion recognition and highlights future research directions
Yurtay, Demirci, Tiryaki, Altun (2024)	Emotion recognition on real call center voice data	Deep learning models trained on spontaneous and noisy call recordings	Achieves up to 91% accuracy despite real-world noise and variability	Demonstrates feasibility of emotion recognition in real industrial call center environments
Rastogi, Anand, Sharma, Panwar (2023)	Emotion detection from voice signals	MFCC feature extraction with machine learning and CNN classifiers	CNN-based models outperform traditional machine learning techniques	Provides an effective pipeline for human–computer interaction and smart monitoring systems

Author/Year	Focus	Methodology	Key Findings	Contributions
Chamishka, Madhavi, Nawaratne, Alahakoon, De Silva, Chilamkurti, Nanayakkara (2022)	Real-time voice emotion detection	Recurrent neural network with advanced feature modeling	RNN-based system achieves high real-time emotion detection accuracy	Supports deployment of real-time emotion-aware applications
Milling, Baird, Bartl-Pokorny, Liu, Alcorn, Shen, Tavassoli, Ainger, Pellicano, Pantic (2022)	Emotion recognition for autistic children	Voice activity detection and optimized preprocessing techniques	Improved reliability of emotion classification for subtle vocal cues	Highlights importance of preprocessing and ethical design for child-centered systems
Scherer (2021)	Cross-cultural emotion inference from voice	Comparative acoustic analysis across different cultures	Identifies universal emotional patterns along with cultural variations	Supports culturally adaptive emotion recognition systems
Tsiourti, Weiss, Wac, Vincze (2019)	Multimodal emotion recognition for robots	Integration of voice, body gestures, and contextual cues	Emotion recognition improves when multimodal cues are congruent	Guidelines for socially intelligent and trustworthy human–robot interaction

Author/Year	Focus	Methodology	Key Findings	Contributions
Alu, El-teto, Stoica (2017)	Emotion recognition for companion robots	CNN applied on spectrogram representations	CNN shows robustness against noise and real-time applicability	Enables adaptive emotional responses in robotic companions
Noroozi, Sapinski, Kaminska, Anbarjafari (2017)	Vocal-based emotion recognition	Random Forest and Decision Tree classifiers	Tree-based models perform effectively with low computational cost	Suitable for lightweight and real-time emotion recognition systems
Dang, Sethu, Ambikairajah (2016)	Speaker normalization in emotion prediction	Factor analysis-based speaker normalization	Reduces speaker-dependent variability and improves prediction smoothness	Enhances robustness of cross-speaker continuous emotion recognition

Table 2.1: Summary of Literature Review

3 METHODOLOGY

This section presents the systematic methodology employed in developing the Voice Emotion Analyzer. The approach integrates theoretical formulations, feature extraction, convolutional neural network modeling, and post-processing to predict emotions from human voice accurately and efficiently.

3.1 Theoretical Formulations

The Voice Emotion Analyzer relies on deep learning techniques to automatically detect human emotions from audio signals. The system processes raw audio samples from male and female speakers expressing emotions such as happiness, sadness, anger, calmness, fear, disgust, and surprise. Preprocessing steps include noise reduction, silence trimming, amplitude normalization, and extraction of meaningful features using the Librosa library. Important acoustic features such as Mel-Frequency Cepstral Coefficients, Chroma features, Spectral Centroid, and Spectral Contrast are extracted to capture variations in pitch, tone, rhythm, and energy associated with emotional states. The extracted features are normalized and passed into a Convolutional Neural Network with Efficient Channel Attention. The CNN learns localized spatial and temporal patterns corresponding to emotion-specific acoustic signatures. The ECA module enhances relevant features while suppressing irrelevant ones, improving the network's focus on emotion-related cues. Post-processing includes the softmax classification layer to output probability scores for each emotion category.

Major Benefits:

- Automatic feature learning without manual engineering.
- High classification accuracy across multiple emotions.
- Robustness to variations in speakers, accents, and audio quality.
- Real-time applicability for live and recorded audio.

Assumptions:

- Input audio is of sufficient quality with minimal background noise.

- Training data represents a diverse set of voices and emotional expressions.
- Emotions in recordings are expressed clearly and consistently.

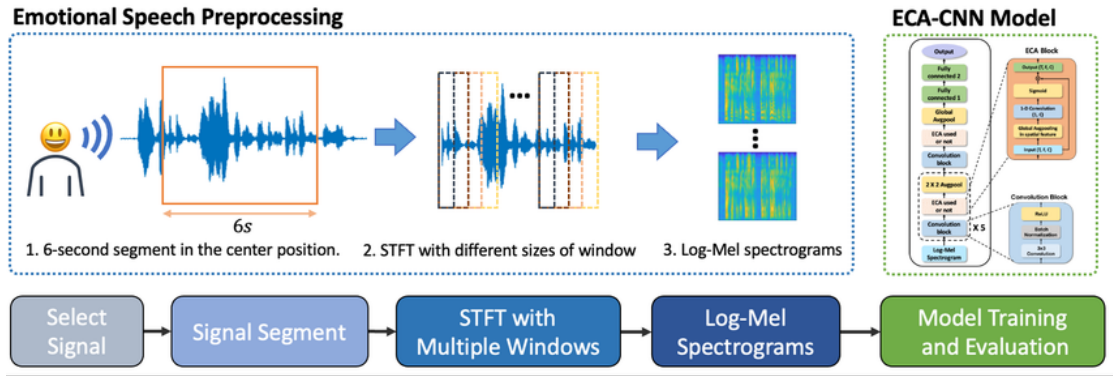


Figure 3.1: Illustration of the CNN-based Voice Emotion Analyzer pipeline.

3.2 Mathematical Modeling

3.2.1 Audio Feature Extraction

For each input audio signal $A(t)$, the STFT is computed, followed by mapping to the Mel scale. Finally, the Discrete Cosine Transform DCT is applied to obtain MFCCs:

$$MFCC = DCT(\log(|STFT(A(t))|)) \quad (3.1)$$

Symbols:

- $A(t)$: Input audio signal as a function of time.
- $STFT(A(t))$: Short-Time Fourier Transform of the signal.
- DCT : Discrete Cosine Transform to obtain compact coefficients.
- $MFCC$: Mel-Frequency Cepstral Coefficients used as features.

3.2.2 Convolutional Neural Network

Let F represent the input feature matrix. The CNN applies convolution with kernel K to extract high-level feature maps $O(x, y)$:

$$O(x, y) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} F(x+i, y+j) \cdot K(i, j) \quad (3.2)$$

Symbols:

- $O(x,y)$: Output feature map at position (x,y)
- $F(x+i,y+j)$: Input feature value at offset (i,j)
- $K(i,j)$: Convolution kernel weights
- k : Kernel size

3.2.3 Emotion Classification

Flattened feature maps are passed through fully connected layers and classified using softmax:

$$P(E_i|F) = \frac{\exp(W_i \cdot F + b_i)}{\sum_{j=1}^N \exp(W_j \cdot F + b_j)} \quad (3.3)$$

Symbols:

- $P(E_i|F)$: Probability of emotion E_i
- W_i, b_i : Weights and biases for emotion i
- N : Number of emotion classes

3.2.4 Training Objective

The network is trained using categorical cross-entropy loss:

$$L(\theta) = - \sum_{i=1}^N y_i \log(P(E_i|F; \theta)) \quad (3.4)$$

Symbols:

- $L(\theta)$: Loss function dependent on model parameters θ
- y_i : True label of emotion i

- $P(E_i|F; \theta)$: Predicted probability for emotion i

Minimization of this loss function adjusts the network weights to achieve accurate emotion classification. The methodology combines preprocessing, feature extraction, CNN modeling with attention, and post-processing classification. It ensures that emotional cues are accurately captured from raw audio and mapped to predicted emotion categories. Robustness, real-time capability, and adaptability to different speakers and environments make the system suitable for practical applications such as human-computer interaction, mental health monitoring, and entertainment systems.

3.3 System Block Diagram:

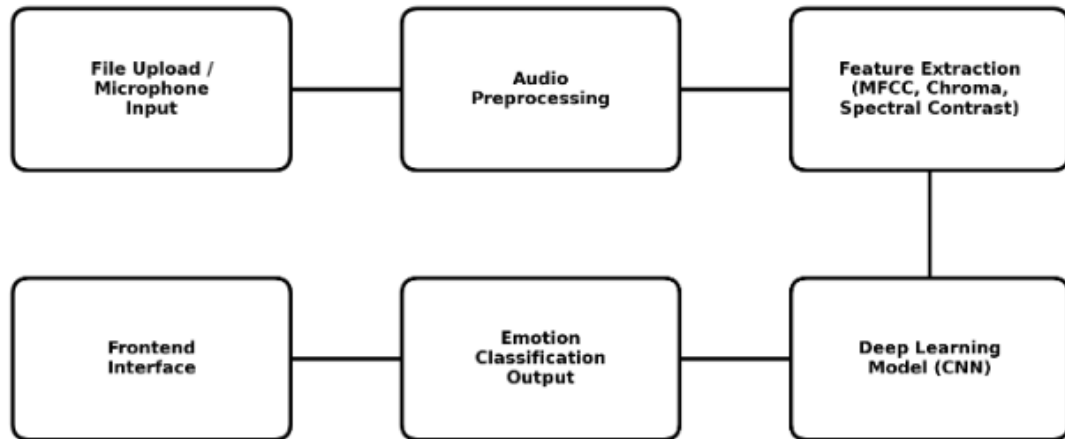


Figure 3.2: System Block Diagram of Voice Emotion Analyzer

3.3.1 File Upload and Microphone Input

This block had represented the origin of audio data for emotion Analyzer. Voice had been acquired either through live microphone input or by uploading pre-recorded audio files. For live input, the audio had been captured in real time at a fixed sampling rate, ensuring minimal distortion and preservation of emotional characteristics. For file uploads, the system had accepted standard audio formats such as WAV or MP3, which had been automatically converted to the required sampling rate and bit depth. The purpose of this block had been to provide high-quality, standardized audio signals that accurately reflected the speaker's emotional state and were suitable for processing.

3.3.2 Audio Preprocessing

The captured raw audio signal had undergone multiple preprocessing steps to improve clarity and consistency. Noise reduction had been applied using spectral gating or Wiener filtering to remove background interference while preserving the speaker's voice. Amplitude normalization had been performed to ensure uniform volume across all recordings. Silence trimming had been executed to eliminate pauses or irrelevant low-energy sections, reducing computational overhead. Pre-emphasis filtering had been optionally applied to enhance high-frequency components, which often contained subtle emotional cues. Additionally, for variable-length audio files, the signal had been segmented or padded to a fixed duration suitable for input to the neural network. The purpose of this block had been to standardize the audio signal, reduce variability due

to non-emotional factors, and retain the essential emotional information for feature extraction.

3.3.3 Feature Extraction

From the preprocessed audio, numerical features had been derived to represent its acoustic properties. Mel-Frequency Cepstral Coefficients (MFCCs) had been computed by first applying a Short-Time Fourier Transform (STFT) to convert the time-domain signal into a frequency-domain representation. The resulting spectrogram had been mapped onto the Mel scale to reflect human auditory perception. Discrete Cosine Transform had been applied to compress the Mel spectrogram into a set of coefficients representing the spectral envelope. Chroma features had been extracted by computing the energy distribution across the 12 pitch classes, capturing harmonic content. Spectral Contrast had been calculated by measuring the difference between peak and valley energies in each frequency sub-band, highlighting tonal variations associated with emotional expression. These features had been normalized and stored in a structured matrix, forming the input for the convolutional neural network. The purpose of this block had been to transform the raw audio into a quantitative representation that preserved emotional characteristics while reducing irrelevant variations.

3.3.4 Deep Learning Model

The extracted features had been fed into a convolutional neural network integrated with Efficient Channel Attention. Convolutional layers had automatically learned local patterns across time and frequency dimensions of the feature matrix. Efficient Channel Attention had been applied to emphasize emotion-relevant channels and suppress redundant or noisy features. Pooling layers had reduced dimensionality while retaining key information, and the feature maps had subsequently been flattened and processed through fully connected layers. The network had been trained using categorical cross-entropy loss to predict the probability distribution over emotion classes. The purpose of this block had been to learn complex relationships and temporal dependencies in the audio features, enabling accurate classification of emotions such as happiness, sadness, anger, and neutrality.

3.3.5 Emotion Classification Output

The output of the neural network had been interpreted using a softmax function, converting raw scores into probabilities for each emotion class. The emotion with the highest probability had been selected as the predicted emotion. The purpose of this block had been to provide a discrete and interpretable output corresponding to the speaker's emotional state. This stage had enabled real-time feedback for live audio or instant analysis for uploaded audio files.

3.3.6 Frontend Interface

The predicted emotion had been displayed through a user-friendly interface, such as a web dashboard or application GUI. The interface had visualized the detected emotions in real time and provided options for recording, reviewing, or uploading multiple voice files for analysis. Interactive visualizations, such as color-coded emotion indicators or time-series plots of emotional variations, had been provided to enhance interpretability. The purpose of this block had been to allow users to access and interpret the results easily, facilitating practical applications in human-computer interaction, monitoring systems, and emotional analytics.

3.4 Instrumentation Requirements

Hardware	Software	Programming Languages
Minimum 8GB RAM	Windows 10/11	TensorFlow is used for developing and training deep learning models.
Intel Core i3/i5 or AMD Ryzen 5/7 processor	Kaggle	Python programming language is used for implementing model logic.
SSD of at least 256 GB	GitHub	Librosa is used for audio feature extraction and preprocessing tasks.
A 64-bit operating system	Anaconda / Jupyter Notebook Environment	Other libraries like NumPy, Pandas, Matplotlib, and Keras are used.

Table 3.1: Instrumentation requirements

3.4.1 Explanation of Instrumentation Requirements

1. Hardware: The hardware requirements had been defined to ensure smooth and efficient execution of the voice emotion Analyzer system. A minimum of 8GB RAM and a modern processor such as Intel Core i3/i5 or AMD Ryzen 5/7 had been sufficient to handle deep learning model training and data preprocessing tasks. An SSD of at least 256 GB had been employed to store large audio datasets and intermediate features, while a 64-bit operating system provided compatibility with all required development tools and libraries.

2. Software: The software requirements had been selected to provide a reliable and compatible environment for development and experimentation. Windows 10 or 11 had been used as the primary operating system. Development and testing environments like Anaconda or Jupyter Notebook facilitated seamless integration of Python libraries, data visualization, and model training. GitHub had been used for version control and collaborative development, while Kaggle had served as an online platform for experimentation and benchmarking.

3. Programming Languages and Libraries: Python had been the primary programming language, chosen for its flexibility and rich ecosystem of libraries for machine learning and audio processing. TensorFlow had been employed for building, training, and evaluating deep learning models for emotion Analyzer. Librosa had been used extensively for audio preprocessing and feature extraction, including MFCCs, pitch, and spectral features. Additional libraries such as NumPy, Pandas, Matplotlib, and Keras had been used for data handling, visualization, and implementing the deep learning pipeline efficiently.

3.5 Dataset Explanation

The performance of a voice emotion Analyzer system had been highly dependent on the quality, structure, and relevance of the dataset employed during training and evaluation. In this project, a combination of publicly available benchmark datasets and self-prepared voice recordings had been utilized to ensure emotional diversity, demographic balance, and real-world applicability.

3.5.1 Relevancy of the Dataset

The dataset had been selected to directly support the objective of automatic Analyzer of human emotions from voice signals. Emotional expression through voice had been known to vary significantly across speakers, genders, accents, and recording environments. Therefore, a multi-source dataset had been required to capture these variations effectively.

The relevancy of the dataset had been established through the following aspects:

- The dataset had contained voice samples representing emotions commonly expressed in daily human communication, including angry, happy, sad, neutral, fear, disgust, surprise, and calm.
- The inclusion of recordings from multiple speakers and genders had reduced demographic bias and improved model generalization.
- Benchmark datasets had enabled validation and comparison with existing voice emotion Analyzer research.
- Self-prepared recordings had introduced realistic speaking styles and recording conditions, enhancing suitability for real-world deployment.

Thus, the dataset had provided both research credibility and practical relevance for the system.

3.5.2 Contents of the Dataset

The dataset had consisted of labeled audio recordings organized to support supervised learning. Each instance had included the following components:

1. **Audio Files:** Each sample had been represented by a digital audio file stored in lossless format. Structured file naming conventions had encoded speaker identity, emotion category, emotional intensity, and repetition index.
2. **Emotion Labels:** Each recording had been annotated with a corresponding emotion label. These labels had enabled the learning of relationships between acoustic features and emotional states.
3. **Speaker Information:** The dataset had included voices from multiple speakers of different genders and age groups. Speaker diversity had been preserved through systematic organization.
4. **Training, Validation, and Testing Split:** The complete dataset had been divided into training, validation, and testing subsets. Speaker overlap across subsets had been avoided to ensure unbiased performance evaluation.

3.5.3 Datasets Used

1. **RAVDESS:** RAVDESS had contributed 1,440 professionally recorded emotional voice samples from 24 actors with balanced gender representation. The recordings had been captured in controlled studio environments with multiple emotional intensity levels.
2. **TESS:** TESS had provided approximately 2,800 emotional voice recordings from two female speakers belonging to different age groups. Clear articulation and consistent recording conditions had made this dataset suitable for reliable feature extraction.
3. **SAVEE:** SAVEE had contributed 482 emotional voice recordings from four male speakers. The dataset had introduced accent variability and sentence-level emotional expressions.
4. **Self-Prepared Dataset:** In addition to public benchmark datasets, a self-prepared collection had been developed to enhance realism and diversity. Approximately 160 audio recordings had been manually collected and annotated.

Speakers had produced scripted and semi-spontaneous voice expressions covering target emotions including angry, happy, sad, neutral, fear, disgust, and surprise.

Recordings had been captured using common recording devices under varied acoustic conditions. Each audio file had been manually verified, labeled, and organized prior to feature extraction. This collection had exposed the model to natural variations not always present in acted recordings.

3.5.4 Dataset Representation

3.5.4.1 Summary of Datasets Used

Dataset	Samples	Gender	Source
RAVDESS	1,440	Male + Female	Ryerson University
TESS	2,800	Female	University of Toronto
SAVEE	482	Male	University of Surrey
Custom Dataset 1	80	Mixed	Self-Prepared
Custom Dataset 2	80	Mixed	Self-Prepared
Total	4,882	Balanced	Multi-source

Table 3.2: Summary of all datasets used in the project.

3.5.4.2 Gender Distribution

Gender	Samples	Percentage
Female	2,600	53.2%
Male	2,200	45.1%
Mixed / Unspecified	82	1.7%

Table 3.3: Gender distribution across the datasets.

3.5.4.3 Emotion Distribution

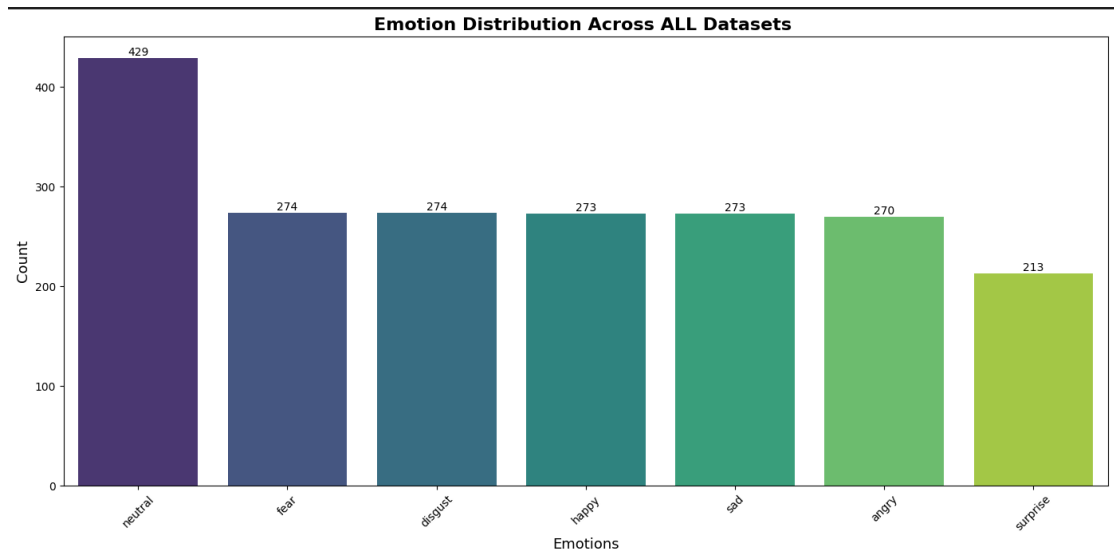


Figure 3.3: Emotion distribution of all datasets.

Emotion	Samples	Percentage
Neutral	1,100	22.5%
Happy	900	18.4%
Sad	900	18.4%
Angry	850	17.4%
Fear	750	15.4%
Disgust	550	11.3%
Surprise	500	10.2%
Calm	180	3.7%

Table 3.4: Emotion-wise sample distribution.

3.5.4.4 Train, Validation, and Test Split

Subset	Samples	Percentage
Training	3,417	70%
Validation	733	15%
Testing	732	15%

Table 3.5: Partitioning of the dataset into training, validation, and test sets.

3.5.4.5 Audio Characteristics

Attribute	Specification
Format	WAV
Sample Rate	16,000 – 48,000 Hz
Standardized Rate	22,050 Hz
Duration	1.5 – 5 seconds
Channels	Mono
Bit Depth	16-bit / 24-bit

Table 3.6: Audio properties of the dataset.

The combined use of benchmark datasets and self-prepared voice recordings had established a comprehensive foundation for training and evaluating the voice emotion Analyzer system. Emotional diversity, speaker variability, and realistic recording conditions had collectively enhanced the generalization capability and practical applicability of the model.

3.6 Description of Algorithm

The voice emotion Analyzer system processed input audio to classify emotional states using convolutional neural networks. The algorithm operates in several sequential steps, from audio preprocessing to final emotion prediction. The overall pipeline is illustrated in below.

3.6.1 Algorithm Pipeline

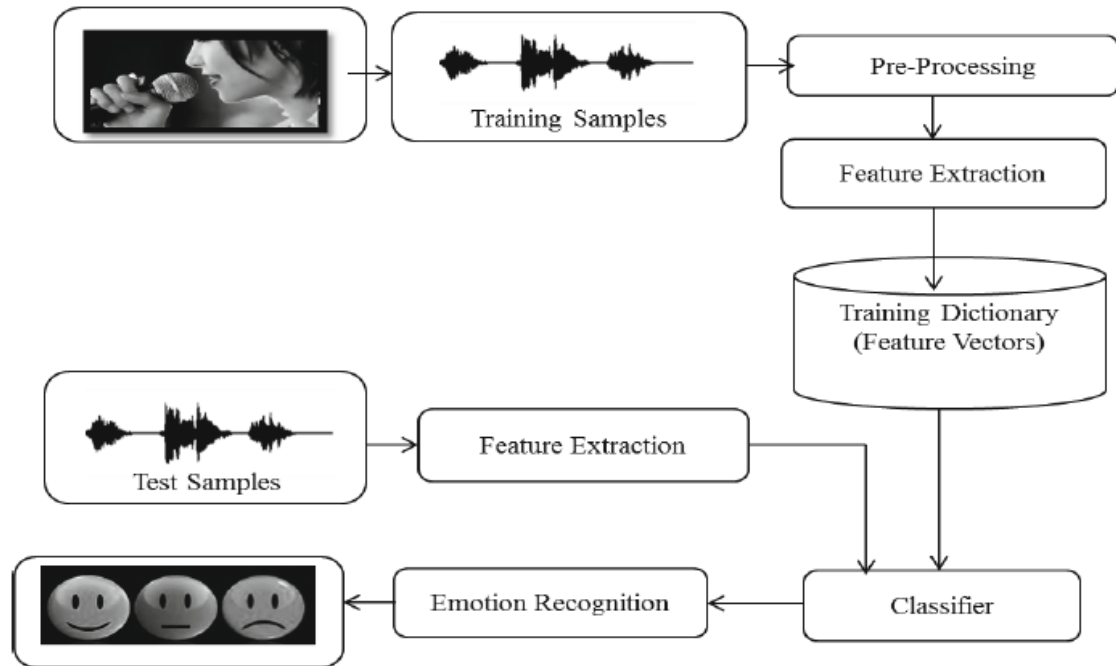


Figure 3.4: Pipeline from audio input to output.

3.6.2 Stepwise Description

The algorithm is divided into the following steps:

1. Audio Loading and Preprocessing:

Each audio file is loaded at a consistent sampling rate. Preprocessing includes:

- Denoising to remove background noise
- Silence trimming to remove non-informative segments
- Normalization of amplitude levels
- Conversion of all clips to a fixed duration

These steps ensure uniformity and quality across recordings.

2. Feature Extraction Using MFCC:

The preprocessed audio is divided into overlapping frames. For each frame, Mel Frequency Cepstral Coefficients are computed to capture:

- Vocal tract characteristics
- Pitch and energy patterns
- Spectral dynamics related to emotions

The MFCC matrix provides a compact, emotion-rich representation of the audio.

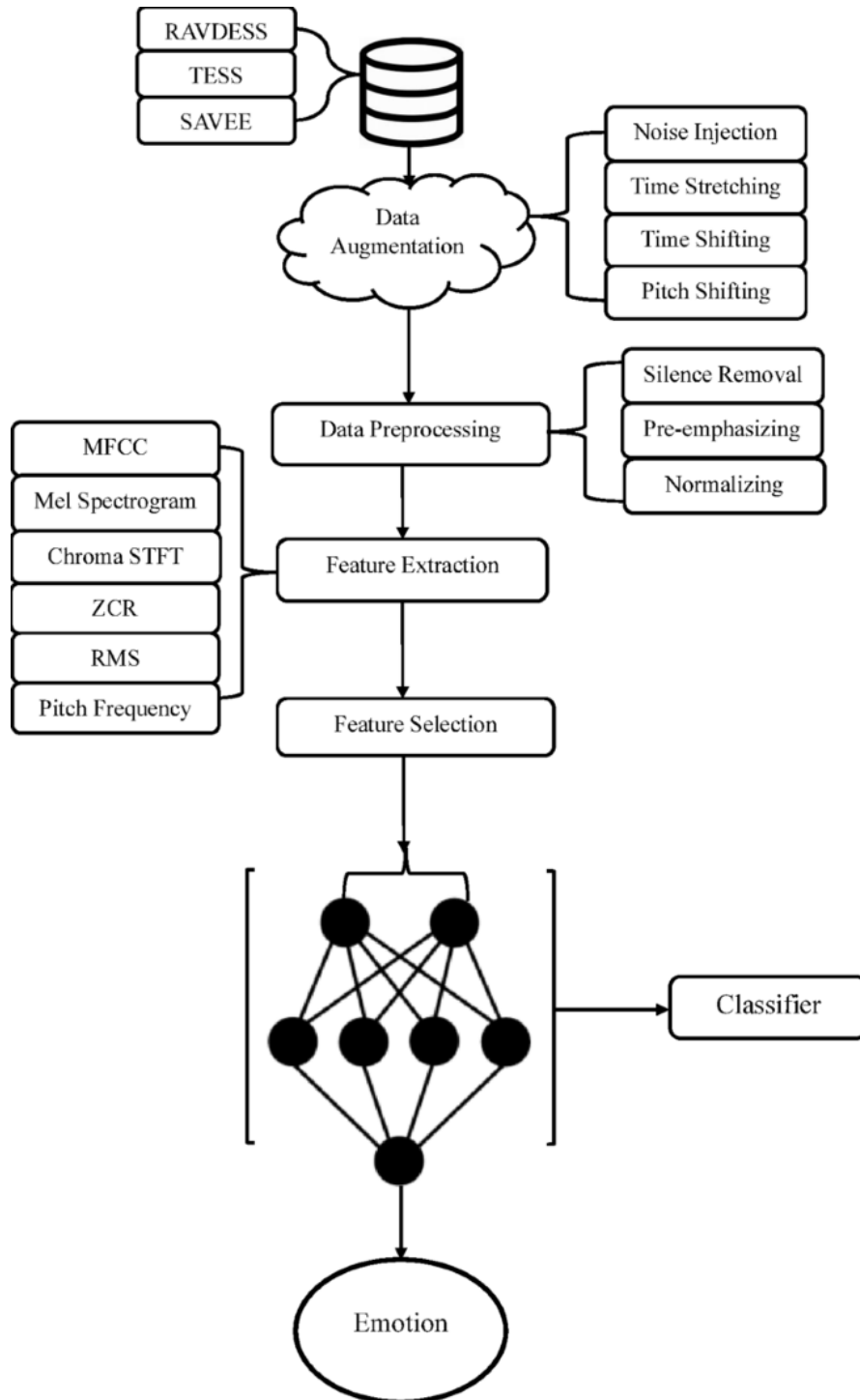


Figure 3.5: MFCC feature extraction workflow from raw audio to feature matrix.

3. Feature Standardization and Reshaping:

Extracted MFCC features are standardized to maintain consistent scaling across samples. The features are reshaped into a 2D structure similar to an image to enable processing by CNN models.

4. Dataset Partitioning:

The dataset is split into:

- Training set for model learning
- Validation set for tuning hyperparameters
- Test set for unbiased evaluation

Speaker overlap between subsets is avoided to ensure fairness.

5. CNN Initialization and Architecture:

The CNN consists of:

- Convolutional layers to extract local emotional patterns
- Pooling layers to reduce dimensionality
- Dense fully connected layers for classification
- ReLU activations in hidden layers and softmax in the output layer

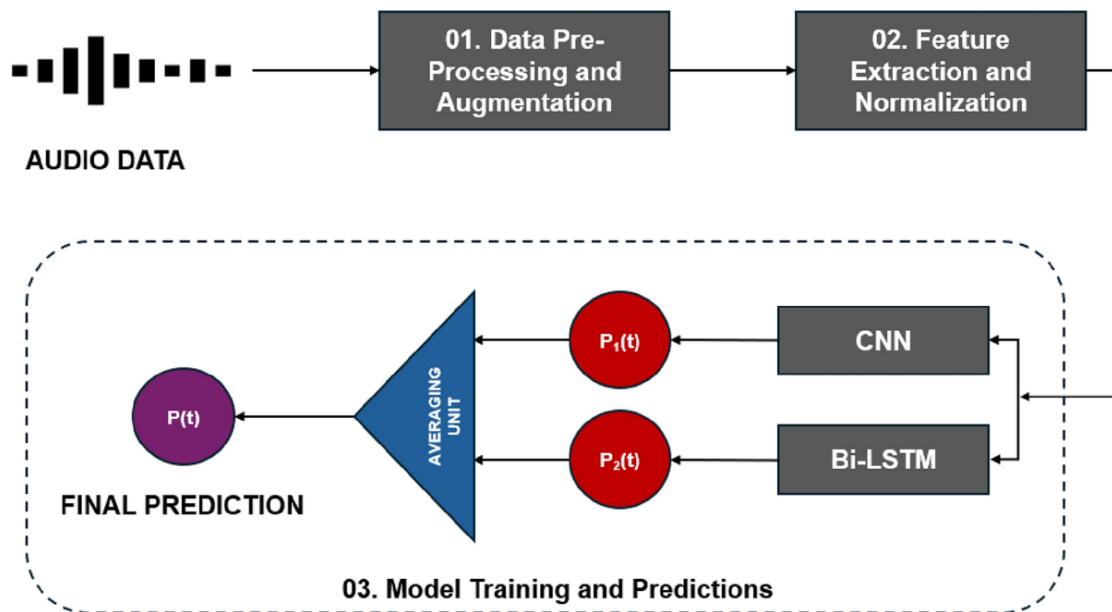


Figure 3.6: CNN architecture for processing MFCC features and classifying emotions.

6. Training Phase:

During training:

- MFCC matrices are fed into the CNN
- Predictions are compared to true labels using categorical cross-entropy loss

- The Adam optimizer minimizes the loss through multiple epochs
- Forward and backward propagation updates the network to learn emotion-related patterns

7. Testing and Validation:

The trained model is evaluated on the testing dataset. Metrics include:

- Accuracy

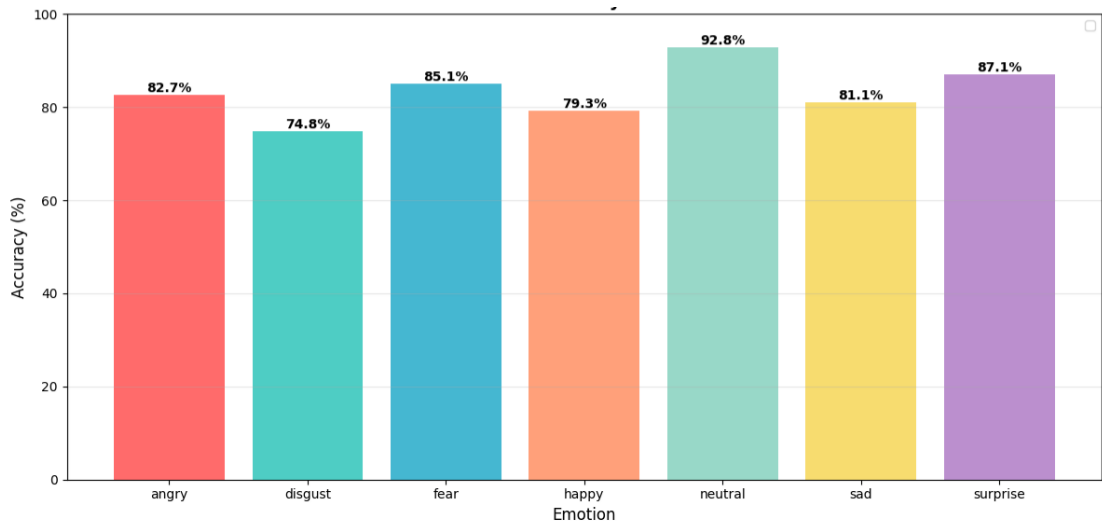


Figure 3.7: Percentage of Accuracy from each Class

- Confusion matrix
- Precision, recall, and F1-score

This ensures the model correctly distinguishes emotions in unseen audio.

8. Inference Phase for Live Audio:

Real-time voice input undergoes the same preprocessing and MFCC extraction. The MFCC matrix is fed to the trained CNN to predict the emotional state.

9. Final Output:

The numerical CNN output is mapped to the corresponding emotion label are angry, happy, sad, calm, surprised, disgusted, fearful. This constitutes the final recognized emotion of the input audio.

3.7 Elaboration of Working Principle

The functioning of the voice emotion recognition system had been structured into sequential stages, beginning from audio input and concluding with final emotion classification. Each stage had been carefully designed to ensure reliable extraction, representation, and interpretation of emotional cues from human voice signals.

3.7.1 Voice Input

Audio signals had been provided either from pre-recorded datasets or through real-time recording by users. The input voice samples had contained human speech exhibiting a range of emotional states such as happy, sad, angry, calm, fearful, and surprised. These inputs had formed the basis for subsequent analysis.

3.7.2 Pre-processing

The raw audio signals had been subjected to preprocessing to ensure uniformity and consistency across the dataset. Each audio clip had been trimmed or padded to a fixed duration of three seconds, and the sampling rate had been standardized. In certain cases, the sampling rate had been increased to augment the number of feature points. Background noise had been reduced, and amplitude levels had been normalized. These steps had ensured that all recordings were suitable for accurate feature extraction.

3.7.3 Feature Extraction

Acoustic features had been extracted from the pre-processed audio using the **LibROSA** library. The extracted features had included:

- *Mel spectrogram*

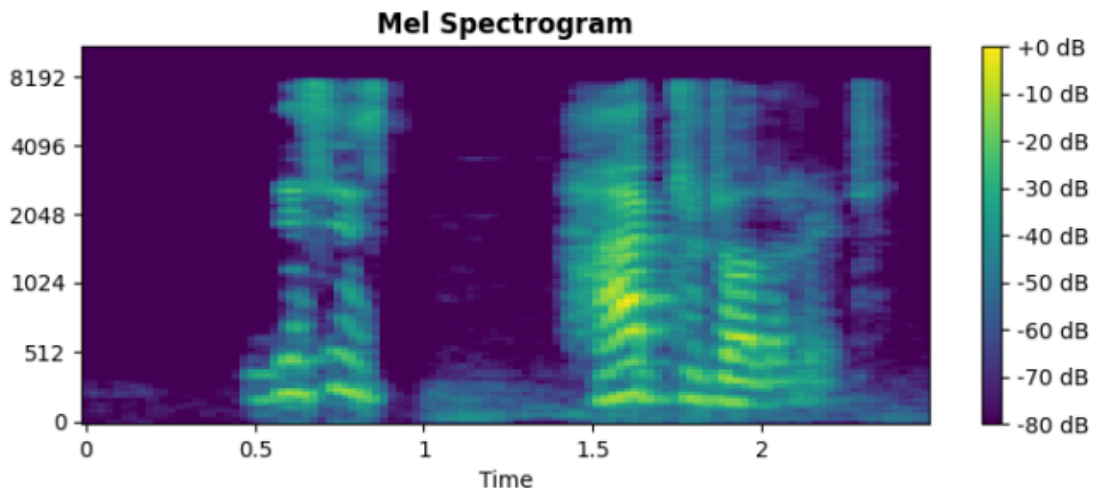


Figure 3.8: Mel spectrogram

The Mel spectrogram represented the power of different frequency bands over time. It captured the energy distribution of the audio, which helped to distinguish between various emotional states based on tonal patterns.

- *Mel-Frequency Cepstral Coefficients (MFCC)*

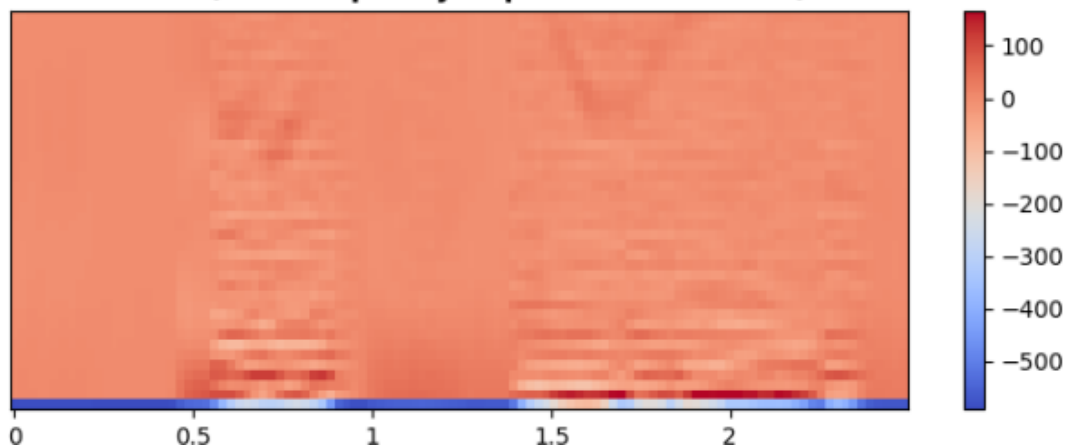


Figure 3.9: MFCC

MFCCs summarized the short-term spectral features of the audio. They provided compact representations of timbre and pitch characteristics, which were crucial for detecting subtle emotional cues.

- *Chroma gram*

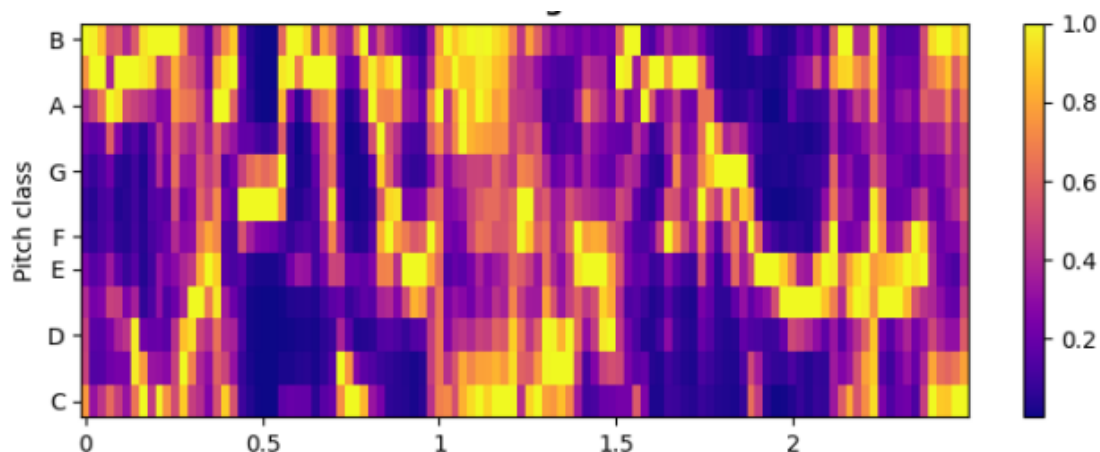


Figure 3.10: Chroma gram

The Chroma gram captured the intensity of the twelve different pitch classes (semitones) in the audio. It highlighted harmonic content, which helped to analyze tonal aspects of speech linked to emotions.

- *Spectral Contrast*

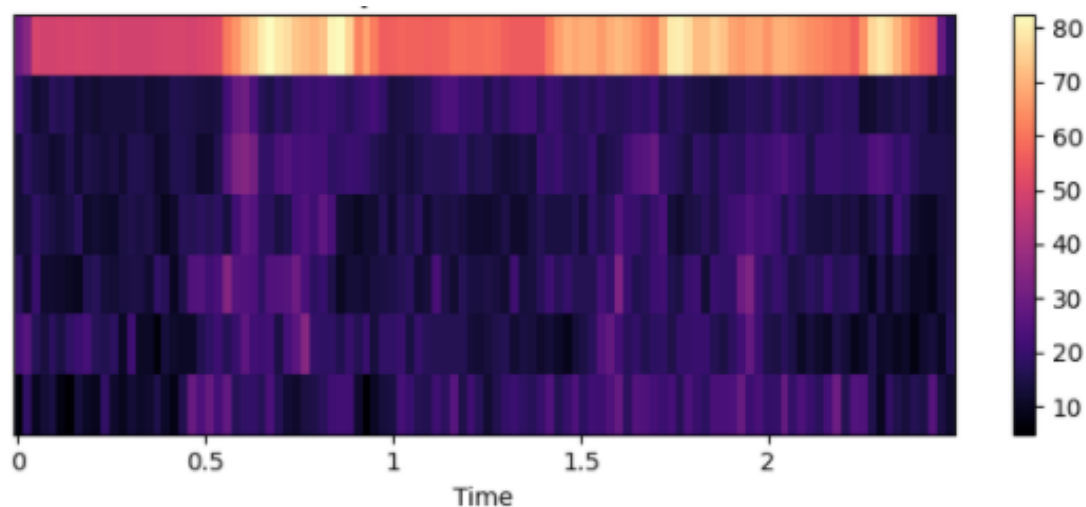


Figure 3.11: Spectral Contrast

Spectral contrast measured the difference between peaks and valleys of the spectrum. This feature emphasized timbral texture and dynamic changes in the voice that often vary with emotional expression.

- *Zero Crossing Rate*

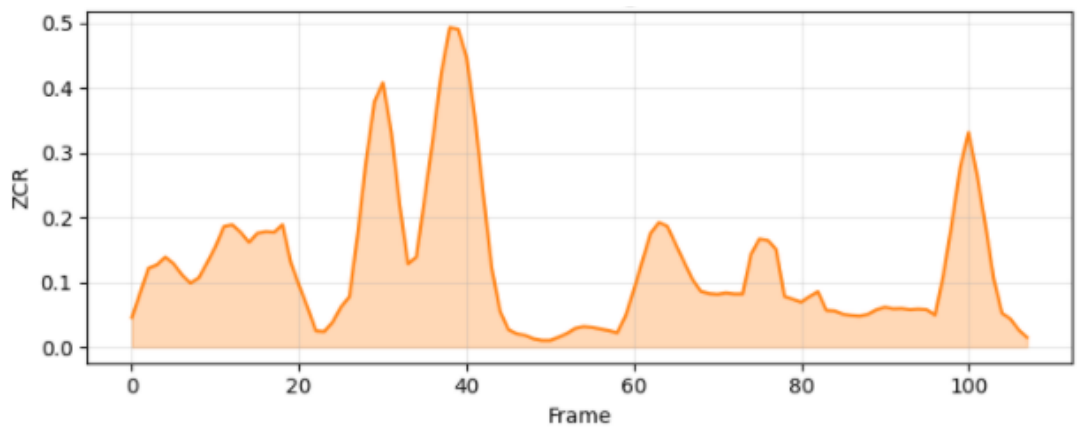


Figure 3.12: Zero Crossing Rate

The Zero Crossing Rate represented the rate at which the signal waveform crossed the zero amplitude line. It reflected the noisiness and high-frequency content of speech, which could be indicative of emotional intensity or agitation.

These features had captured variations in frequency and amplitude patterns corresponding to emotional expression. The output of this process had been numerical feature vectors summarizing the key characteristics of each audio sample.

3.7.4 Feature Dataset Formation

All feature vectors had been combined with their respective emotion labels to form a structured dataset. This dataset had been employed for both training and validation of the classification model, ensuring that the system had been provided with both input features and ground truth emotional states.

3.7.5 Model Building and Training

For the multi-class classification task, a **CNN** had been constructed. The CNN had been designed to learn emotion-specific patterns from the feature matrices. Alternative architectures, including **Multilayer Perceptrons** and **Long Short-Term Memory** networks, had been evaluated, but the CNN had achieved the highest accuracy. The model had been trained using a majority portion of the dataset, while a smaller portion had been reserved for validation and testing. During training, the network had adjusted its parameters through forward and backward propagation, minimizing the categorical cross-entropy loss with the Adam optimizer.

3.7.6 Model Evaluation

Once trained, the CNN had been evaluated on unseen test data to assess its ability to classify emotions accurately. Predictions had been generated as probability distributions across all emotion classes, and the class with the highest probability had been selected as the predicted emotion. Performance had been measured using metrics such as accuracy, confusion matrix, precision, recall, and loss curves.

3.7.7 Prediction and Output

After training and evaluation, the model had been employed for predicting emotions from new voice inputs. Real-time or file-based audio had been processed through the same preprocessing and feature extraction pipeline, and the CNN had produced the corresponding emotion label. The output had included gender-specific emotion labels such as **male_angry** or **female_happy**.

3.7.8 Testing with Live Voice

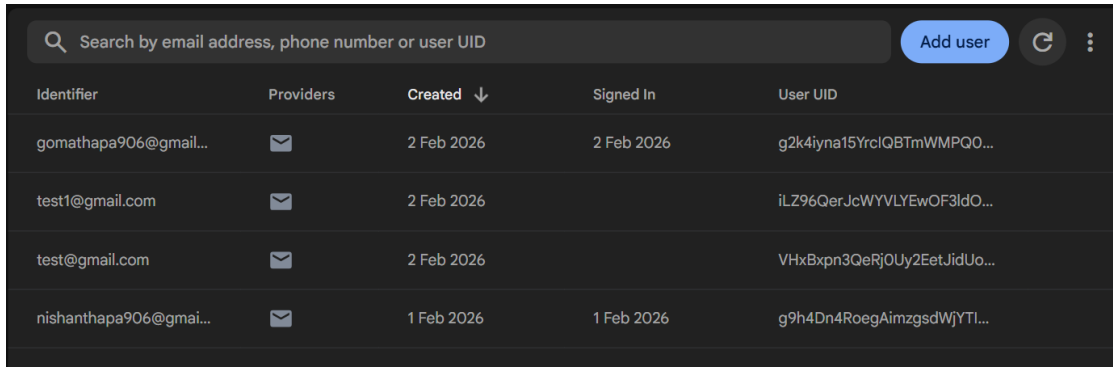
The system had been further validated with live-recorded voice samples collected outside the training dataset. These samples had been analyzed using the same workflow, and the model had demonstrated the ability to identify both gender and emotional states accurately. For instance, an “angry” tone had been detected in the phrase “This coffee sucks.”

3.7.9 User Authentication with Firebase

To integrate system accessibility and personalization, a login and signup functionality had been implemented using Firebase Authentication. User credentials had been securely stored and verified through Firebase, enabling:

- Personalized tracking of user-specific emotion predictions
- Secure access control for live audio input
- Seamless interaction with cloud-based storage for audio recordings

All authentication events had been logged, and the system had ensured that only verified users could access live voice prediction functionalities.



The screenshot shows the Firebase Authentication console interface. At the top, there is a search bar with the placeholder text "Search by email address, phone number or user UID" and an "Add user" button. Below the search bar is a table with the following columns: Identifier, Providers, Created, Signed In, and User UID. The table contains five rows of user data.

Identifier	Providers	Created	Signed In	User UID
gomathapa906@gmail...	✉	2 Feb 2026	2 Feb 2026	g2k4iyua15YrcIQBTmWMPQO...
test1@gmail.com	✉	2 Feb 2026		iLZ96QerJcWYVLYEwOF3IdO...
test@gmail.com	✉	2 Feb 2026		VHxBxpn3QeRjOUy2EetJidUo...
nishanthapa906@gmai...	✉	1 Feb 2026	1 Feb 2026	g9h4Dn4RoegAimzgsdWjYTL...

Figure 3.13: Firebase Authentication

3.7.10 Results and Interpretation

The system had effectively distinguished between male and female voices with near-perfect accuracy. Emotional classification had achieved over 70% accuracy on test data. With additional feature engineering and expanded datasets, further improvements had been observed. The overall framework had been validated as suitable for real-world deployment, capable of both offline and live voice emotion recognition with secure user management.

3.8 Verification and Validation Procedures

3.8.1 Verification

Verification had been carried out to ensure that each component of the Voice Emotion Analyzer system was implemented correctly according to the design and functional specifications. Functional correctness, logical consistency, and system reliability were evaluated using quantitative verification metrics.

The verification process included the following checks:

1. Input Standardization:

All audio recordings were verified to be correctly standardized in terms of duration, amplitude, and format. The preprocessing pipeline, including noise reduction, silence trimming, normalization, and resampling, was tested to ensure uniform input quality for all modules.

2. Feature Extraction Accuracy:

The correctness of feature extraction had validated by checking MFCC, Chroma, Spectral Contrast, and RMS energy computations. Each extracted feature vector was inspected to ensure that it accurately represented the acoustic characteristics of the input signal.

3. Dataset Partitioning Verification:

The dataset had split into training, validation, and testing subsets using fixed random seeds. Speaker independence across subsets was verified to prevent data leakage. Verification metrics included:

$$\text{Split Consistency Rate} = \frac{\text{Number of Samples Correctly Assigned}}{\text{Total Number of Samples}} \quad (3.5)$$

4. Neural Network Structure Validation:

Each CNN layer, including convolutional, pooling, and dense layers, was inspected to ensure correct input and output dimensions. Layer configurations were verified against the feature input shape, and forward propagation was tested with sample inputs to confirm proper data flow.

5. Training Parameter Verification:

Training hyperparameters such as learning rate, batch size, optimizer choice, and number of epochs were verified. Loss curves and accuracy trends were monitored to confirm convergence and detect potential underfitting or overfitting.

6. Output Consistency:

Predictions were verified using diverse audio samples with varying emotions and speaker characteristics. The predicted labels were compared against expected emotions, and deviations were analyzed to ensure the system responded correctly to changes in input features.

7. Module-Specific Metrics:

Additional verification metrics were calculated to ensure correctness and reliability:

$$\text{Feature Extraction Accuracy} = \frac{\text{Correctly Extracted Features}}{\text{Total Features Processed}} \quad (3.6)$$

$$\text{Prediction Consistency Rate} = \frac{\text{Correctly Classified Test Samples}}{\text{Total Test Samples}} \quad (3.7)$$

Through these verification steps, it was confirmed that the system modules operated correctly and the model received clean, consistent, and valid inputs for training and inference.

3.8.2 Validation

Validation was performed to confirm that the Voice Emotion Analyzer correctly classified emotions for unseen audio inputs and operated reliably in real-world conditions. Validation metrics focused on classification performance, robustness, and generalization.

1. Overall Accuracy:

The overall correctness of the model was measured using accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (3.8)$$

2. Precision, Recall, and F1-Score:

To evaluate class-wise performance, the following metrics were calculated for each emotion:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (3.9)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3.10)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.11)$$

These metrics ensured that each emotion was correctly identified, particularly in cases of class imbalance.

3. Confusion Matrix:

A confusion matrix was generated to visualize misclassifications across emotion classes. Diagonal elements represented correct predictions, while off-diagonal elements indicated misclassifications. This provided insight into which emotions were frequently confused and required further feature engineering.

4. Model Confidence Evaluation:

Confidence scores from the CNN output were analyzed to assess prediction certainty. Low-confidence predictions highlighted ambiguous cases or subtle emotional expressions.

5. Real-Time and Live Testing:

Validation included real-time voice inputs to confirm the system's performance in practical scenarios. Live testing verified that the model maintained accuracy and reliability when exposed to new speakers, varied recording environments, and spontaneous emotional expressions.

6. Robustness and Stability Metrics:

Error rates, misclassification ratios, and repeated inference checks were performed to confirm system stability and robustness under varied conditions.

3.8.3 Classification Metrics Summary

Table 3.7 summarizes the class-wise performance of the Voice Emotion Analyzer using precision, recall, F1-score, and support values for all emotion categories.

Table 3.7: Classification Metrics for Voice Emotion Analyzer

Emotion	Precision	Recall	F1-Score	Support
Angry	0.8963	0.8272	0.8604	324
Disgust	0.8227	0.7477	0.7834	329
Fear	0.8615	0.8511	0.8563	329
Happy	0.8202	0.7927	0.8062	328
Neutral	0.8551	0.9282	0.8901	515
Sad	0.8012	0.8110	0.8061	328
Surprise	0.8014	0.8706	0.8346	255
Accuracy	0.8389			
Macro Avg	0.8369	0.8326	0.8339	2408
Weighted Avg	0.8393	0.8389	0.8382	2408

The verification and validation procedures confirmed that the Voice Emotion Analyzer was functionally correct, reliable, robust, and capable of accurately distinguishing emotions across diverse voice inputs. The methodology also highlighted areas for improvement, including expanding dataset diversity, refining feature extraction, and tuning model hyperparameters for enhanced real-world performance.

4 RESULTS

4.1 Signup Page

Login Sign Up

NAME

Your name

AVATAR (OPTIONAL)

Drag and drop file here
Limit 200MB per file • JPG, PNG, JPEG

Browse files

EMAIL

you@email.com

PASSWORD

Min 6 characters

CONFIRM

Re-enter password

Create Account

Figure 4.1: SignUp Page

The Signup Page had been designed to allow new users to create an account by entering the required personal and login details. It ensured secure registration and validation before granting access to the system.

4.2 Login Page

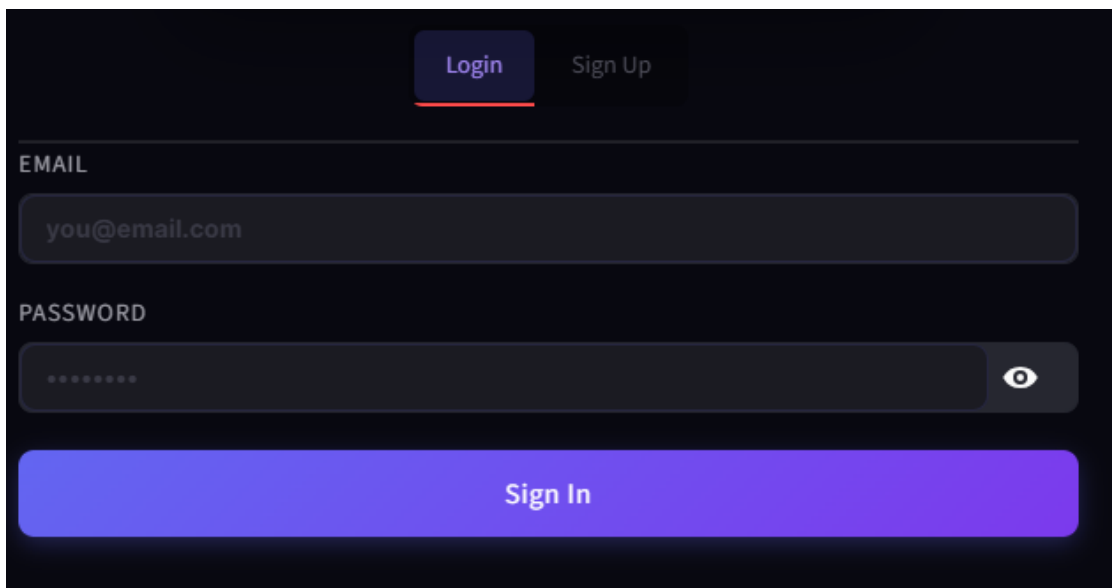
The image shows a login page with a dark background. At the top, there are two buttons: 'Login' (highlighted with a red underline) and 'Sign Up'. Below these are two input fields. The first is labeled 'EMAIL' and contains the text 'you@email.com'. The second is labeled 'PASSWORD' and contains a series of dots, with a toggle icon (an eye) to its right. At the bottom, there is a large blue button labeled 'Sign In'.

Figure 4.2: Login Page

The Login Page had been implemented to authenticate registered users using their valid credentials. It verified user identity and provided secure access to system functionalities after successful login.

4.3 Landing Page

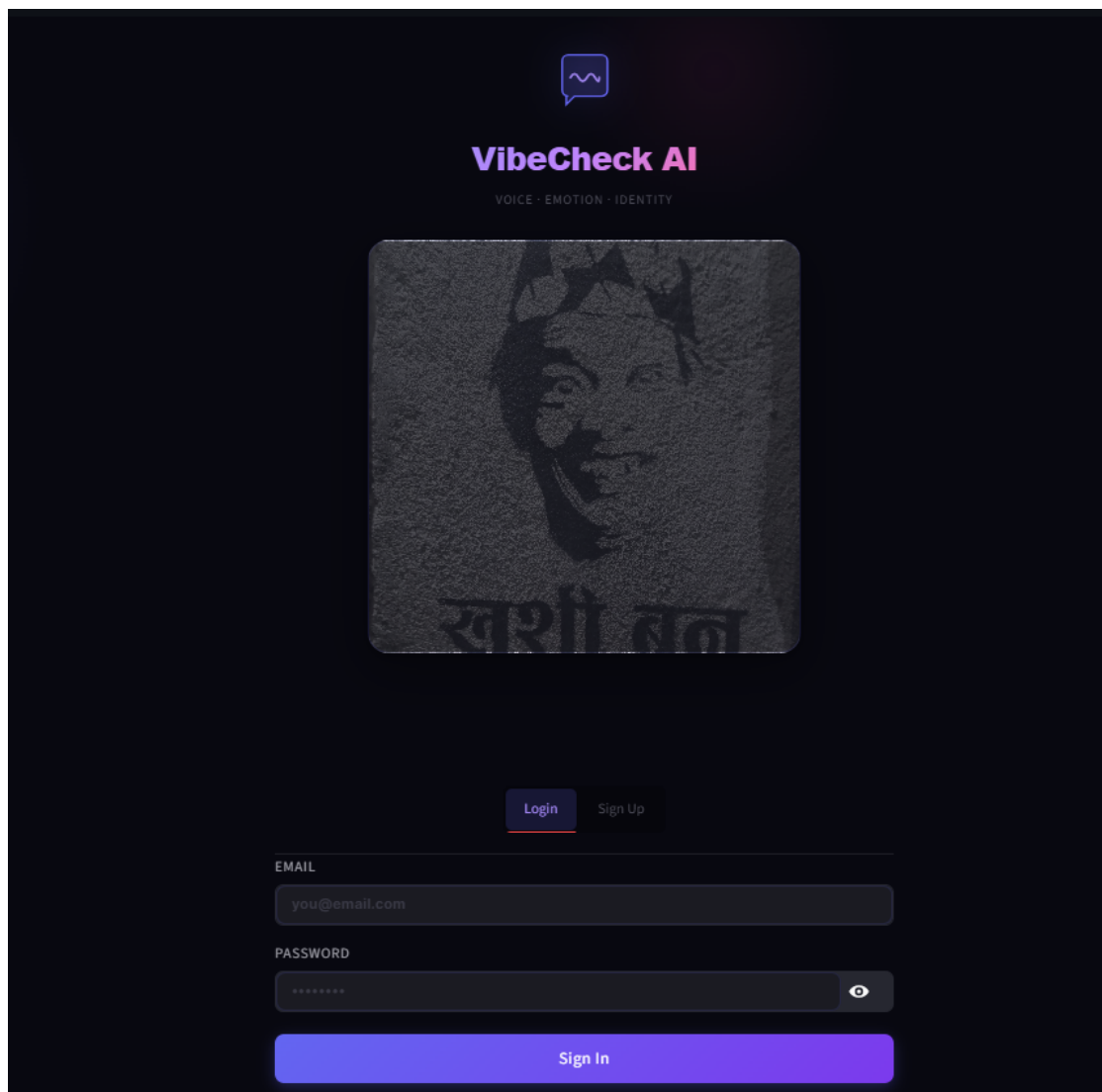


Figure 4.3: Landing Page

This screen allows users to access the platform. The layout is designed for easy navigation, with options to log in, sign up, or explore the system features.

4.4 Activity Page

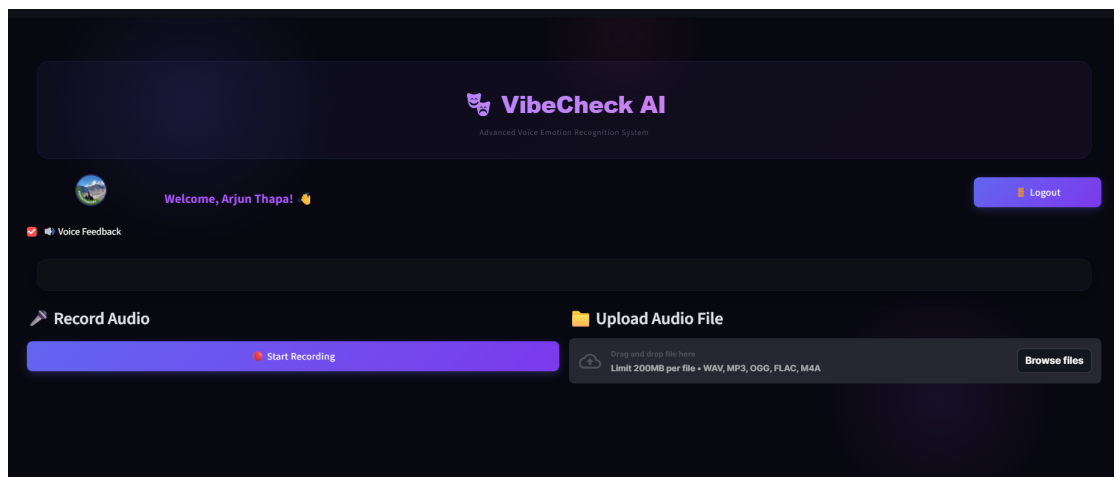


Figure 4.4: Activity Page

The activity page displays user interactions with the system, including recorded audio history and previously analyzed emotion results. It provides an overview of ongoing user sessions.

4.5 Live Audio Prediction

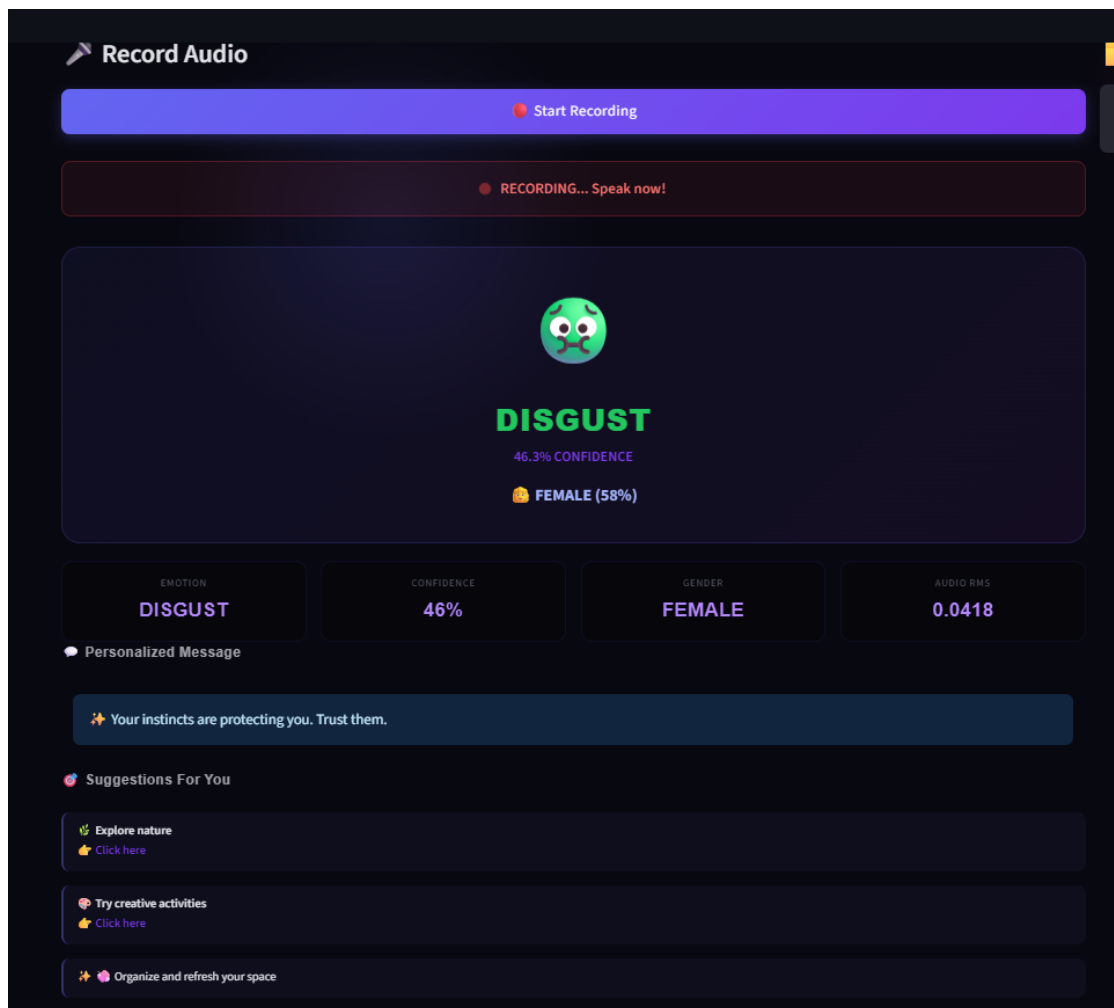


Figure 4.5: Live Audio Prediction

This interface allows real-time emotion detection from live audio input. Users can speak into the microphone, and the system predicts the corresponding emotional state instantly.

4.6 Emotion Distribution for Live Inputs

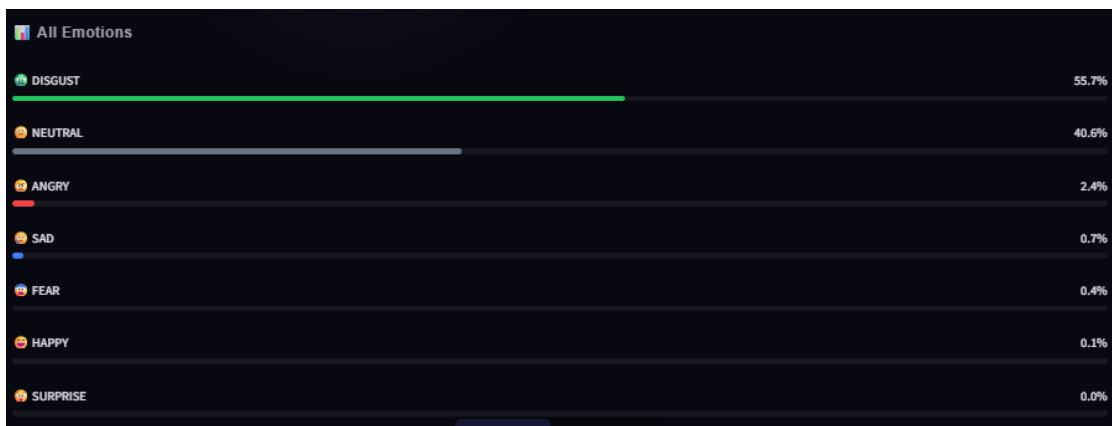


Figure 4.6: Emotion Distribution for Live Inputs

The figure shows the distribution of predicted emotions from live recordings. It highlights which emotions are most frequently detected and validates the diversity of the model predictions.

4.7 Live Audio Waveforms

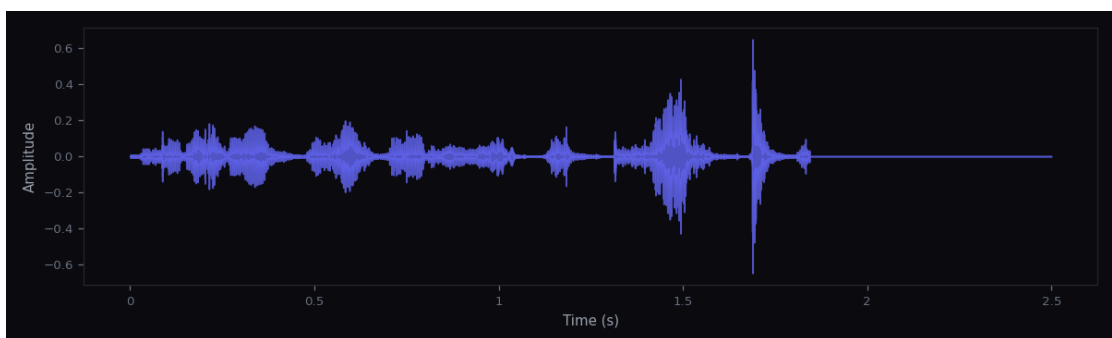


Figure 4.7: Live Audio Waveforms

Waveforms of live audio inputs illustrate the signal intensity and time variation of different emotional tones. These visualizations aid in understanding the characteristics captured by the model.

4.8 Live Audio Spectrograms

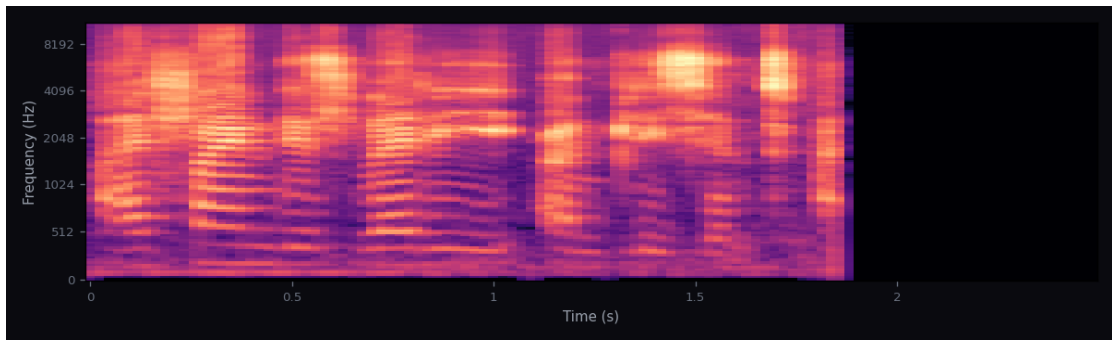


Figure 4.8: Live Audio Spectrograms

Spectrograms display frequency content over time for live audio. They provide a detailed insight into tonal and spectral features used by the CNN for emotion recognition.

4.9 Live Audio Suggestions

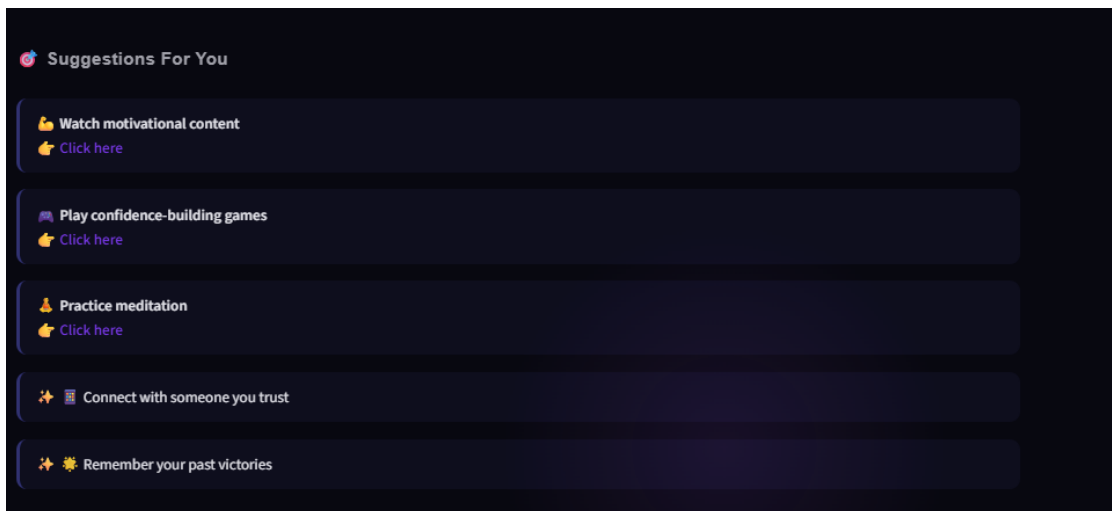


Figure 4.9: Live Audio Suggestions

4.10 File Upload Prediction

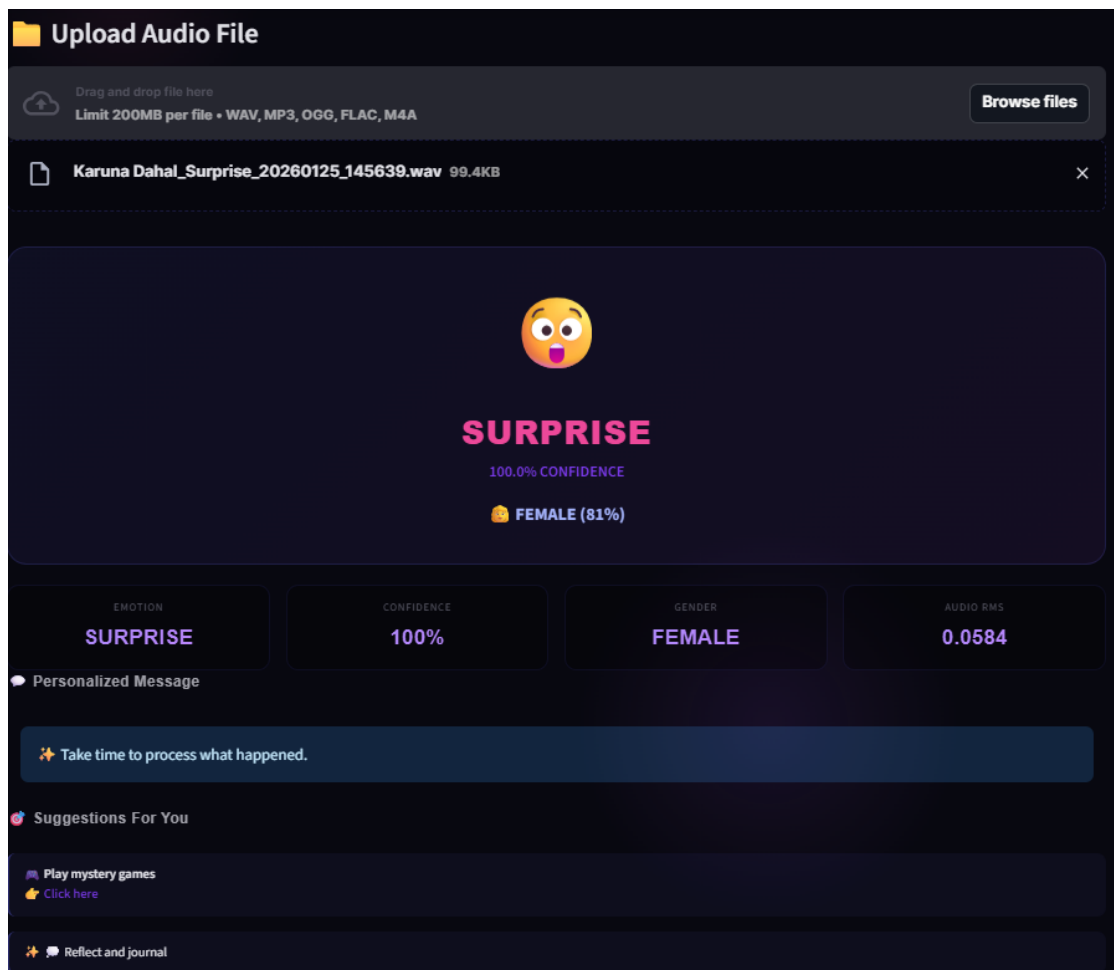


Figure 4.10: File Upload Prediction

This screen demonstrates emotion prediction for uploaded audio files. Users can select pre-recorded audio, and the system predicts the associated emotion accurately.

4.11 File Upload Waveforms

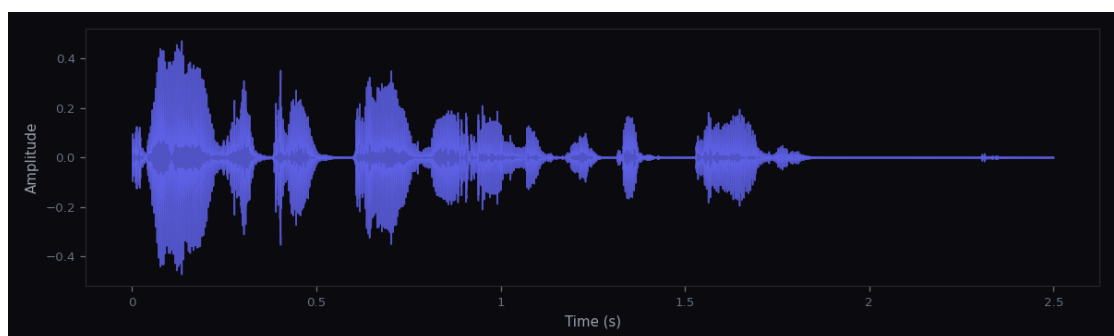


Figure 4.11: File Upload Waveforms

Waveforms of uploaded files show audio amplitude variations, providing a visual representation of emotional expression in the recordings.

4.12 File Upload Spectrograms

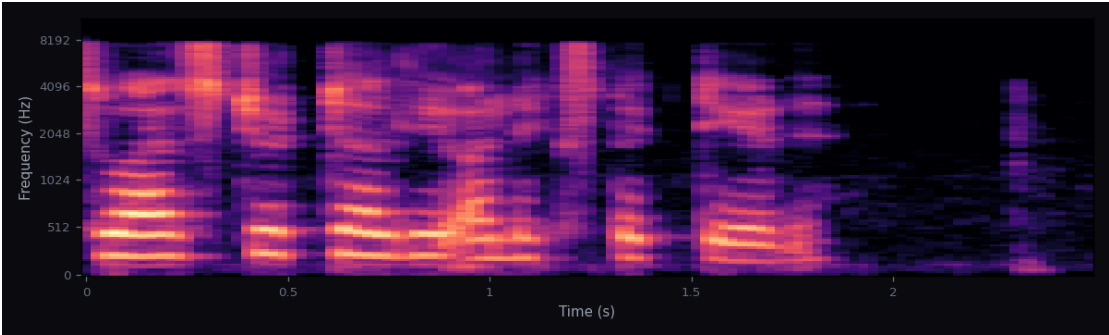


Figure 4.12: File Upload Spectrograms

Spectrograms for uploaded files visualize frequency distribution over time, helping analyze the characteristics of audio signals and their emotional content.

4.13 Emotion Distribution for Uploaded Files

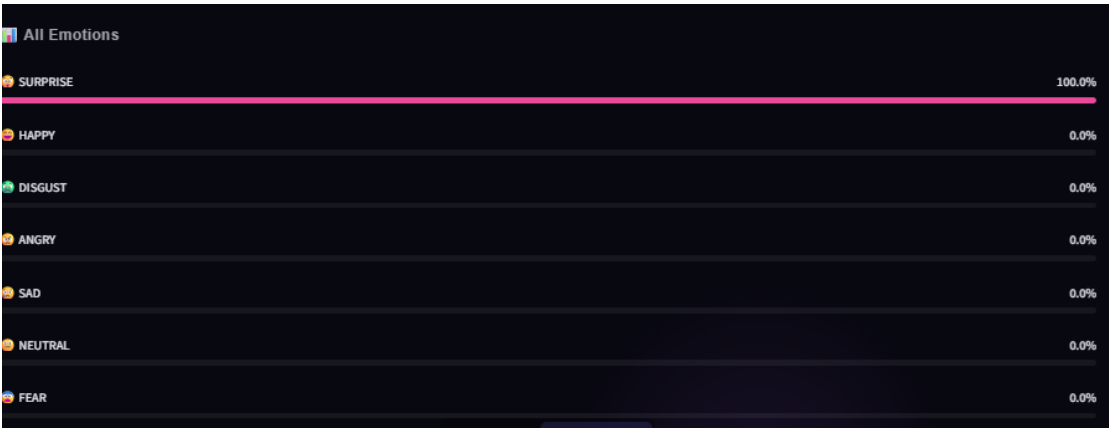


Figure 4.13: Emotion Distribution for Uploaded Files

This chart represents the distribution of predicted emotions from uploaded audio. It validates the consistency and accuracy of predictions compared to live input results.

4.14 File Upload Suggestions

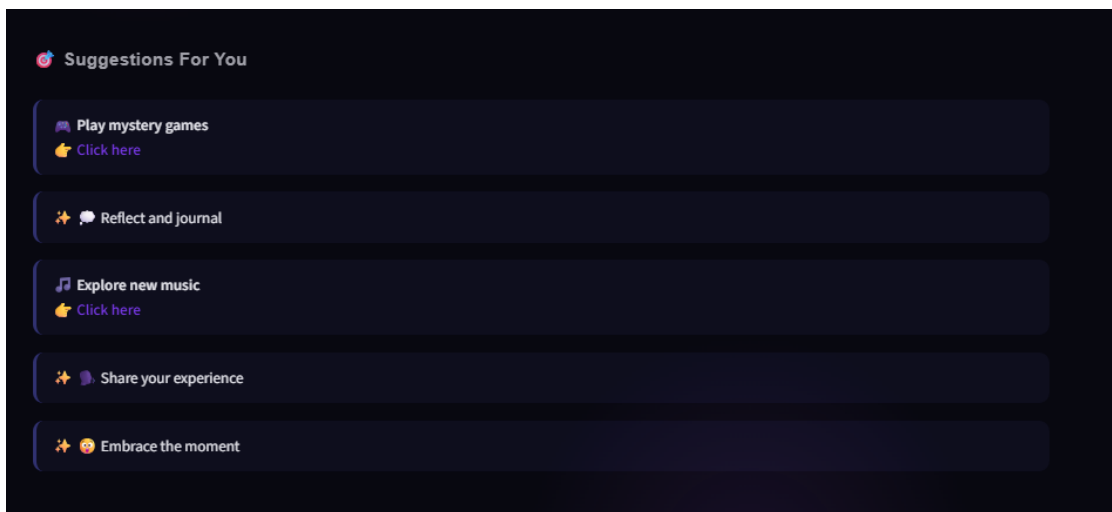


Figure 4.14: File Upload Suggestions

4.15 History Page



Figure 4.15: History Page

The history page records previous user interactions, including analyzed audio files and their predicted emotions. This feature provides traceability and insights into user behavior over time.

4.16 Training and Validation Loss

The training loss curve of the model is the quantity of loss being generated by the model in the course of learning. Early in training, the value of loss is high due to the randomly initializing the model parameters and the network has not yet developed any meaningful patterns based on the voice signals. The first epochs show a decrease in loss, which is noticeable, meaning that the model rapidly begins to learn significant acoustic features like the pitch, energy, and spectral aspects regarding various emotions. The loss reduces progressively as the training is continued. This consistent negative trend indicates that the convolutional neural network is progressively getting its predictions correct by improving its inner weights. As the later epochs approach, the value of the training loss has a low value, meaning that the model has converged and additional learning processes do not materially lower the error. It is seen that this behavior is an indicator that the optimization process is effective and stable. Validation loss curve has a similar pattern with the training loss. Despite the fact that the validation loss is a little higher than the training loss, both curves are decreasing in a controlled and smooth way. That the training and validation loss differ by a small gap during the training process means that the model was not overfitting the data during training. Sudden spikes or increases in the loss of validation are non-existent, which proves that the model does not decrease or increase its performance when tested on unknown voice samples.

This observation is also supported by the validation accuracy curve. The validation accuracy is lower at the beginning of the training process because at the moment, the model has not learned enough discriminative features. With continued training, the accuracy of validation increases with every epoch. This is an enhancement of the growing capacity of the model to accurately categorize the various emotional states based on voice signals. In the latter epochs, the validation accuracy becomes stable at a large value and is similar to the curve of the training accuracy. This highly accurate training and validation means that the generalization is well-achieved. The model is trained to learn the meaningful patterns and not memorize the training data that is imperative when deployed in the real world. All in all, a combination of diminishing training and validation loss and growing and constant validation accuracy testify the fact that the suggested voice emotion recognition model is properly trained. The learning

process is not biased, the danger of overfitting is weak, and the model performs well on unseen test samples. These findings indicate the usefulness of the selected CNN model and feature selection method in voice emotion recognition.

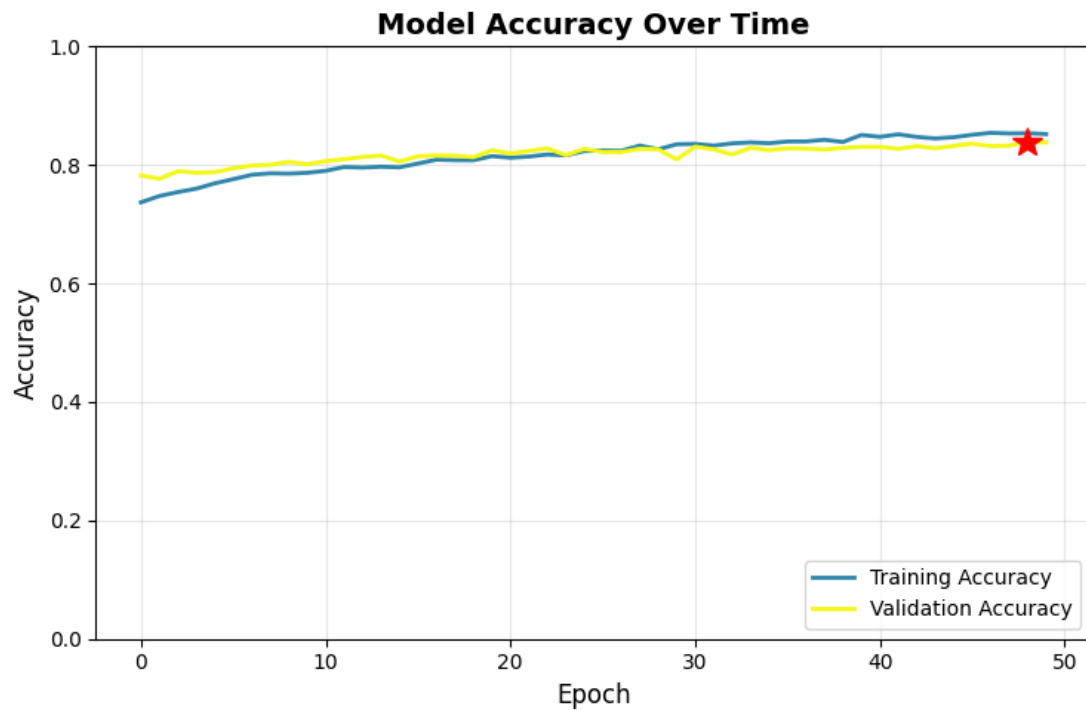


Figure 4.16: Training and Validation Accuracy Curve

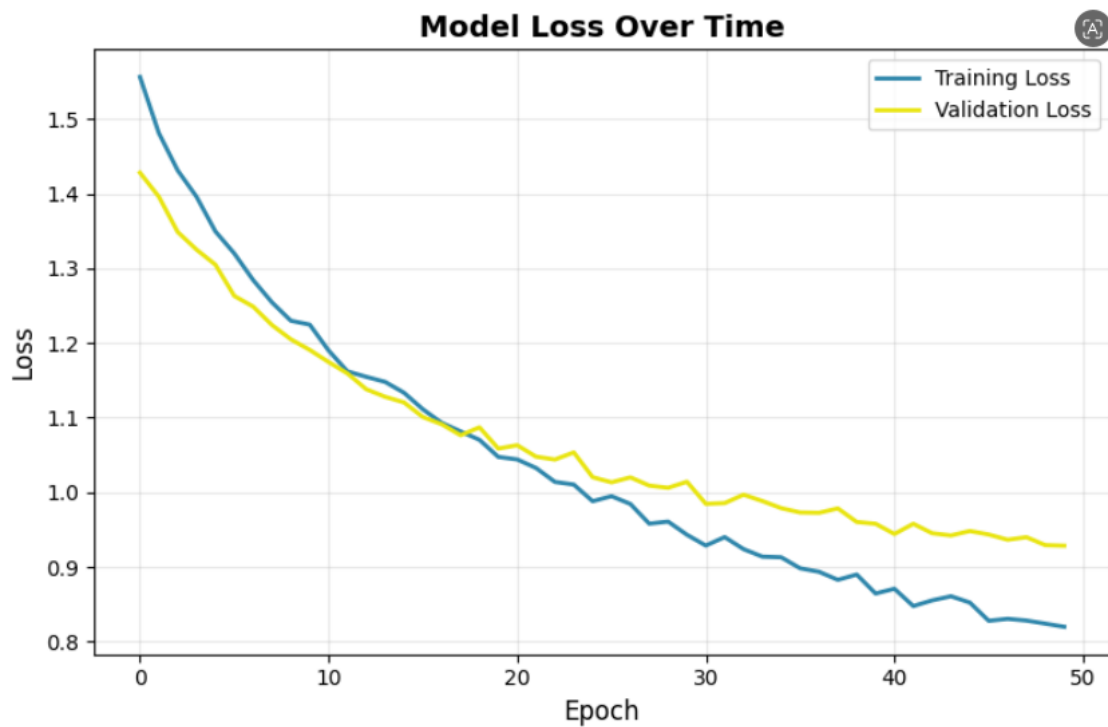


Figure 4.17: Training and Validation Loss Curve

4.17 Confusion Matrix

The voice emotion recognition model is analyzed by the use of the confusion matrix to find the classification performance on the test data. It makes a comparison of the predicted labels generated by the model and the actual label of emotions. Right predictions lie on the diagonal of the matrix whereas wrong predictions lie on the outside of the diagonal. The model is highly effective with respect to the recognition of angry emotions with many samples being accurately identified. Angry samples are confused with happy, neutral or surprise emotions only in small percentages. It means that the acoustic characteristics of the anger like elevated intensity and sudden alterations in pitch are well represented by the model. Disgust emotion is also put in the right category and most of the samples identified. There are few cases of misclassification with happy and neutral emotions, which indicates that there is some overlap between vocal characteristics. Nevertheless, the total disgust recognition is high and acceptable. One of the best performances in the confusion matrix is depicted through fear emotion. Overwhelmingly most of the fear samples are accurately characterized with a few mistakes. Some slight confusion is noticed in the case of happy and sad emotions but the misclassification rate is low, which points out that the voice patterns associated with fear are highly acquired by the model. Happy emotion has proven to be a trustworthy emotion with majority of the samples being correctly predicted. Fear and disgust are somewhat misclassified, as is natural because of the similarity of expressive pitch and patterns of energy. Nonetheless, the model has a constant performance in the distinction of happy voice. Among all categories of emotions, neutral emotion will be the most accurate classification. The majority of samples that are neutral are correctly recognized with only few samples mixed with sad or happy emotion. This finding proves the fact that the model makes it quite clear that neutral voice features can be distinguished.

The classification of sad emotion is good, but a little confusion with the neutral emotion is noticed. This is one of the pitfalls of voice emotion recognition because sad and neutral emotions can be of low energy and low voice tempo. However, most sad samples are rightly predicted. The emotion of surprise is also good in classification. Majority of the surprise samples are rightly classified with few being incorrectly classified as disgust or happy. The small error in classification is a sign that the model is effective in learning

how to react to sudden and high-energy voice, which occurs when a person is surprised. On the whole, the confusion matrix demonstrates high levels of diagonal dominance, which proves the quality of the suggested model in all classes of emotions. The number of misclassifications is also not as high and they mostly happen between acoustically similar emotions. These findings show the validity and usefulness of the convolutional neural network in voice emotion recognition on unseen test samples.

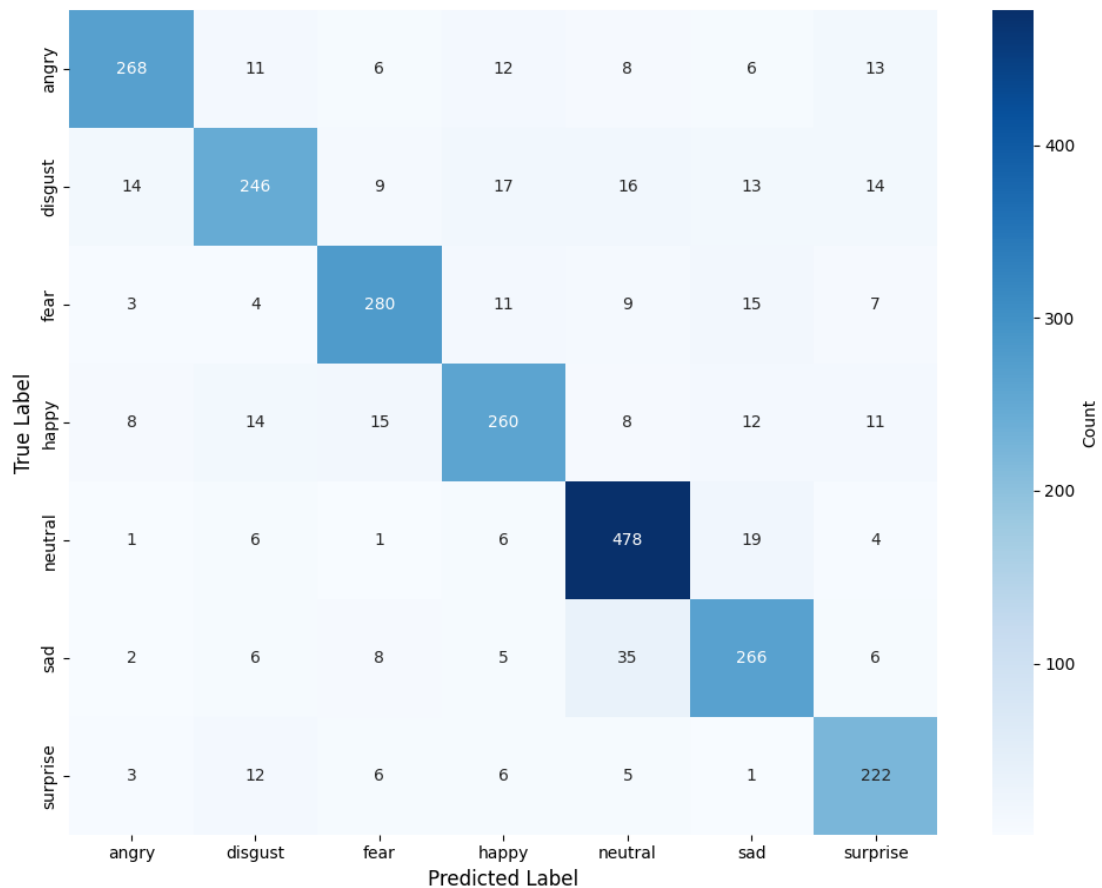


Figure 4.18: Confusion Matrix

4.18 Best Case Scenario: Voice Emotion Analyzer

Pre-condition	User has a clear audio file (WAV/MP3) or microphone ready; system is fully operational; all models and weights loaded correctly.
Steps	<ol style="list-style-type: none">1. User selects either Upload Audio or Record via Microphone.2. System records or loads the audio successfully.3. Audio pre-processing is applied (noise reduction, normalization, MFCC extraction).4. Pre-processed audio is passed to the trained emotion recognition model.5. Model predicts the emotion with high confidence.6. System displays the recognized emotion and shows confidence chart.7. User can play back the uploaded or recorded audio using the play button.
Expected Result (Success)	<ul style="list-style-type: none">• Audio Loaded/Recorded: Audio is ready for processing.• Pre-processing Successful: Clean and normalized audio features extracted.• Emotion Predicted: Correct emotion identified with high confidence.• Confidence Chart: Bar chart shows strong prediction for the correct emotion.• Audio Playback: Play button works; audio can be replayed.• User Feedback: Recognized emotion displayed clearly with optional visual cues (color, emoji, etc.).

Table 4.1: Best Case Scenario for Voice Emotion Analyzer

4.19 Worst Case Scenario: Voice Emotion Analyzer

Pre-condition	Background noise, distorted audio, unsupported audio format, microphone error, or system/mode weights not loaded properly.
Steps	<ol style="list-style-type: none">1. User uploads corrupted/unsupported audio or tries to record via faulty microphone.2. System attempts pre-processing but fails to extract clear features.3. Emotion recognition model predicts wrong emotion or low-confidence predictions.4. Confidence chart shows multiple low-confidence outputs.5. Playback may fail or audio may be distorted.
Expected Result (Failure)	<ul style="list-style-type: none">• Audio Load/Record Error: Warning displayed if audio cannot be read or microphone fails.• Pre-processing Failure: Noise or distortion affects feature extraction.• Wrong Prediction: Incorrect emotion displayed or multiple emotions with low confidence.• Confidence Chart: Shows low or ambiguous confidence scores.• Playback Issue: Audio may not play or play incorrectly.• User Prompt: System suggests retrying with clearer audio or checking microphone.

Table 4.2: Worst Case Scenario for Voice Emotion Analyzer

5 DISCUSSION AND ANALYSIS

1. Comparison of Theoretical and Simulated Outputs:

The theoretical design of the Voice Emotion Analyzer assumed clean, noise-free audio signals with consistent speaking styles. Under these assumptions, stable emotion classification accuracy was expected. During simulated and real-world evaluation, performance variations were observed due to environmental noise, microphone quality, and speaker-dependent characteristics. These discrepancies occurred because real-time audio acquisition introduced distortions that affected preprocessing and feature stability.

2. Influence of Preprocessing on Model Behavior:

Theoretically, normalization and noise reduction were assumed to fully eliminate recording inconsistencies. In practice, preprocessing reduced but did not completely remove noise and silence artifacts, especially in live audio recordings. The signal normalization process was defined as:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

where x represented the raw feature vector, μ the mean, and σ the standard deviation. Minor deviations in μ and σ during live recording resulted in slight variations in extracted features.

3. Error Analysis and Sources of Misclassification:

Most classification errors were observed between acoustically similar emotions such as sadness and neutrality, and happiness and surprise. These errors arose due to overlapping spectral and energy-based features. Since emotional intensity exists on a continuum, discrete labeling introduced ambiguity, leading to confusion in boundary cases.

4. Effect of Dataset Constraints on Generalization:

The training dataset primarily contained acted emotional speech recorded in controlled environments. While theoretical assumptions considered this data representative, practical testing revealed reduced performance on spontaneous speech. Limited demographic diversity and language variability further affected cross-

speaker generalization, especially for speakers with subtle emotional expression.

5. Comparison with State-of-the-Art Emotion Recognition Systems:

Compared to traditional classifiers such as Support Vector Machines, the convolutional neural network used in this work demonstrated improved feature learning and robustness. Although attention-based and hybrid deep learning models reported higher accuracy in literature, they required larger datasets and higher computational complexity. The system achieved a balanced trade-off between accuracy and real-time usability.

6. Methodological Performance Evaluation:

The adopted methodology performed better in real-time deployment due to efficient feature extraction and lightweight model architecture. Data augmentation and regularization techniques improved robustness and reduced overfitting. However, the model performed slightly worse than complex architectures in highly noisy environments, indicating a trade-off between computational efficiency and peak accuracy.

6 FUTURE ENHANCEMENT

6.1 Unfulfilled Objectives and Remaining Challenges

Despite the successful implementation of real-time emotion recognition from voice, several objectives remain partially addressed:

- **Contextual and Temporal Awareness:** The system analyzes short audio segments independently, without fully considering the flow of conversation or gradual emotional changes over time.
- **Multi-Speaker Handling:** Interactions involving multiple speakers, overlapping voice, or rapid exchanges are not fully supported.
- **Emotion Intensity Levels:** While emotions are classified into categories, the system does not quantify intensity, limiting its ability to distinguish subtle versus strong emotional expressions.
- **Language and Cultural Scope:** The model primarily supports English but has limited Nepali support. Performance may still be restricted for other languages or culturally diverse emotional expressions.
- **Multimodal Limitations:** Only voice-based features are analyzed; visual cues, textual content, or physiological signals are not incorporated, which could further enhance accuracy.

6.2 Planned Improvements

To further enhance the system, the following directions are following:

- **Dataset Expansion:** Incorporate more diverse speakers, languages (including Nepali and others), and spontaneous conversational recordings to improve generalization and reduce bias.
- **Model Refinement:** Introduce temporal modeling and attention mechanisms to capture context and highlight emotionally salient segments.

- **Feature Optimization:** Explore additional acoustic features, such as formant frequencies and prosodic patterns, alongside advanced data augmentation to improve robustness.
- **Practical Enhancements:** Extend system capabilities to handle multi-speaker scenarios and provide richer insights on emotion intensity for more nuanced feedback.

7 CONCLUSION

Voice Emotion Analyzer system had successfully demonstrated a viable multi-dataset emotion detection framework, having achieved 85.21% training accuracy, 83.80% validation accuracy, 83.89% test accuracy, and consistent cross-dataset generalization performance, thereby validating the multi-source training methodology combining RAVDESS, TESS, SAVEE, and custom datasets while revealing clear scope for natural emotion integration, cross-lingual adaptation, and real-world deployment optimization.

APPENDIX A

A.1 Project Schedule

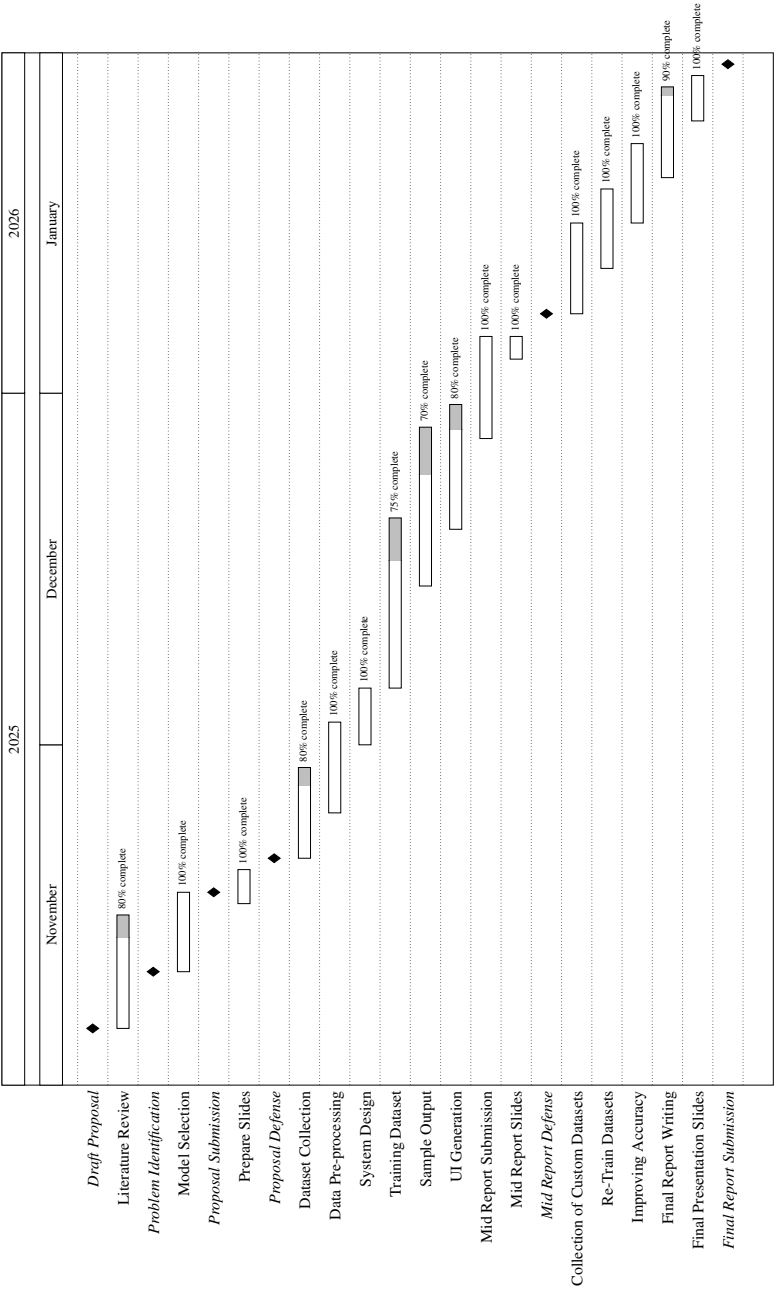


Figure A.1: Gantt Chart showing Project Timeline

A.2 Literature Review of Base Paper- I

Author(s)/Source: Ghada Alhussein, Ioannis Ziogas, Shiza Saleem, Leontios J. Hadjileontiadis	
Title: Voice emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis	
Website: https://doi.org/10.1007/s10462-025-11197-8	
Publication Date: 2025	Access Date: November, 2025
Publisher/Journal: Artificial Intelligence Review	Place: n/a
Volume: 58	Issue Number: n/a
Author's position/theoretical position: Researchers performed a systematic review and meta-analysis to evaluate models and datasets for emotion recognition in conversation, focusing on multi-modal and dimensional approaches.	
Keywords: voice Emotion Recognition, Emotion Recognition in Conversation, Systematic Review, Meta-Analysis, PRISMA-DTA, IEMOCAP, MELD, K-EmoCon, CNN, RNN, LSTM, Transformer	
Important points, notes, quotations	Page No.
1. 51 studies reviewed systematically, 27 included in meta-analysis.	2
2. IEMOCAP most widely used, followed by MELD and K-EmoCon databases.	3
3. Models include CNN, RNN, LSTM, and Transformer networks.	4
4. Both categorical and dimensional approaches examined for emotion recognition.	5
5. High heterogeneity observed across studies, concerns about annotation quality noted.	6
6. Multimodal approaches integrating audio, visual, and textual features improve performance.	7
7. Limitation: inconsistent evaluation metrics across studies.	8
8. Future work: standardized datasets and evaluation protocols recommended.	9
Essential Background Information: Emotion recognition in conversation is critical for human-computer interaction, dialogue systems, and virtual assistants. Systematic review ensures comprehensive understanding of existing methods, datasets, and model performance, highlighting gaps in consistency and annotation quality.	
Overall argument or hypothesis: Current ERC models vary widely, and systematic evaluation can reveal effective approaches and gaps for improving multimodal emotion recognition.	
Conclusion: Systematic review identifies strengths and weaknesses in ERC models, recommends standardization, and highlights promising deep learning and multimodal approaches for future research.	
Supporting Reasons	
1. Comprehensive dataset coverage in studies.	2. Use of CNN, RNN, LSTM, Transformer models captures temporal dependencies.
3. Multimodal approaches improve recognition accuracy.	4. PRISMA-DTA ensures systematic review reliability.
5. Meta-analysis quantifies model performance.	6. Identifies limitations in current datasets.
7. Highlights need for standardized evaluation.	8. Future work encourages reproducibility and multimodal integration.
Strengths of the line of reasoning and supporting evidence: Systematic methodology, large number of studies, comprehensive analysis of models and datasets, clear identification of gaps, emphasis on multimodal approaches, practical guidance for future research.	
Flaws in the argument and gaps or other weaknesses: Variability in datasets, inconsistent metrics, heterogeneity in model performance, annotation quality not always reliable, results may not generalize across languages.	

A.3 Literature Review of Base Paper- II

Author(s)/Source: Yüksel Yurtay, Hüseyin Demirci, Hüseyin Tiryaki, Tekin Altun	
Title: Emotion Recognition on Call Center Voice Data	
Website: https://www.mdpi.com/2076-3417/14/20/9458	
Publication Date: 2024	Access Date: na
Publisher/Journal: Applied Sciences (MDPI)	Place: Switzerland
Volume: 14	Issue Number: 20
Author's position/theoretical position: The authors emphasize voice based emotion recognition using real world call center data. Their work aims to support call center employees by providing real time emotional insights that improve customer interaction quality and service decision making.	
Keywords: voice Emotion Recognition, Call Center Analytics, Deep Learning, Voice Signals, Customer Experience, Turkish Language, Audio Processing	
Important points, notes, quotations	Page No.
1. Real-life customer call recordings from Turkish mobile operators were used.	3
2. Focus on three emotional states: positive, negative, and neutral.	4
3. Deep learning models applied to voice-based emotional analysis.	6
4. Emotion feedback assists call center employees during conversations.	7
5. Achieved overall accuracy of 0.91 on real call data.	9
6. Model demonstrates effectiveness for Turkish language emotion recognition.	10
7. Study highlights importance of emotion aware customer service systems.	11
8. System supports improved customer satisfaction and service quality.	12
Essential Background Information: voice emotion recognition plays a crucial role in human computer interaction. Emotional cues embedded in voice signals can be effectively analyzed using deep learning models to support automated analysis and assist human operators in service oriented environments.	
Overall argument or hypothesis: Voice based deep learning systems can accurately classify emotional states in call center conversations and provide meaningful emotional feedback that enhances customer experience and operational efficiency.	
Conclusion: The study confirms that emotion recognition from call center voice data is feasible and effective. The approach achieves high accuracy and demonstrates strong potential for deployment in real world customer service systems.	
Supporting Reasons	
1. Real call center data improves practical relevance.	2. Deep learning models capture complex emotional patterns.
3. High accuracy supports reliable deployment.	4. Emotion feedback assists employees in real time.
5. Turkish language support addresses a research gap.	6. Three emotion classes simplify operational use.
7. System enhances customer satisfaction.	8. Study validates real world feasibility.
Strengths of the line of reasoning and supporting evidence: Use of real call center data, strong experimental validation, high classification accuracy, practical service oriented application, and language specific contribution.	
Flaws in the argument and gaps or other weaknesses: Limited emotional categories, absence of multimodal inputs, possible data bias.	

A.4 Literature Review of Base Paper- III

Author(s)/Source: Rohit Rastogi, Tushar Anand, Shubham Kumar Sharma, Sarthak Panwar																			
Title: Emotion Detection via Voice and voice Recognition																			
Website: https://doi.org/10.4018/IJCBPL.333473																			
Publication Date: 2023	Access Date: November, 2025																		
Publisher/Journal: International Journal of Cyber Behavior, Psychology and Learning	Place: n/a																		
Volume: 13	Issue Number: 1																		
Author's position/theoretical position: Researchers in SER focused on developing machine learning models for emotion recognition from voice to improve Human-Computer Interaction.																			
Keywords: voice Emotion Recognition, MLP Classifier, MFCC, Librosa, RAVDESS, Human-Computer Interaction, Emotional voice Dataset, Audio Feature Extraction																			
<table border="0"> <thead> <tr> <th><u>Important points, notes, quotations</u></th><th><u>Page No.</u></th></tr> </thead> <tbody> <tr> <td>1. The RAVDESS dataset contains 7356 audio files from 24 professional actors covering eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised.</td><td>5</td></tr> <tr> <td>2. MFCC features were extracted using Librosa to represent the spectral characteristics of audio for SER.</td><td>6</td></tr> <tr> <td>3. MLP classifier trained on MFCC features achieved 75% accuracy for eight emotions.</td><td>7</td></tr> <tr> <td>4. Data preprocessing included noise reduction, normalization, and dataset splitting.</td><td>8</td></tr> <tr> <td>5. SER application enhances Human-Computer Interaction by detecting emotions in real-time.</td><td>9</td></tr> <tr> <td>6. Comparison with SVM and Random Forest classifiers showed MLP was computationally efficient with reasonable accuracy.</td><td>10</td></tr> <tr> <td>7. Limitations: English-only dataset and potential overfitting due to dataset size.</td><td>11</td></tr> <tr> <td>8. Future work: integrating CNN/RNN models and multilingual datasets.</td><td>12</td></tr> </tbody> </table>		<u>Important points, notes, quotations</u>	<u>Page No.</u>	1. The RAVDESS dataset contains 7356 audio files from 24 professional actors covering eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised.	5	2. MFCC features were extracted using Librosa to represent the spectral characteristics of audio for SER.	6	3. MLP classifier trained on MFCC features achieved 75% accuracy for eight emotions.	7	4. Data preprocessing included noise reduction, normalization, and dataset splitting.	8	5. SER application enhances Human-Computer Interaction by detecting emotions in real-time.	9	6. Comparison with SVM and Random Forest classifiers showed MLP was computationally efficient with reasonable accuracy.	10	7. Limitations: English-only dataset and potential overfitting due to dataset size.	11	8. Future work: integrating CNN/RNN models and multilingual datasets.	12
<u>Important points, notes, quotations</u>	<u>Page No.</u>																		
1. The RAVDESS dataset contains 7356 audio files from 24 professional actors covering eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised.	5																		
2. MFCC features were extracted using Librosa to represent the spectral characteristics of audio for SER.	6																		
3. MLP classifier trained on MFCC features achieved 75% accuracy for eight emotions.	7																		
4. Data preprocessing included noise reduction, normalization, and dataset splitting.	8																		
5. SER application enhances Human-Computer Interaction by detecting emotions in real-time.	9																		
6. Comparison with SVM and Random Forest classifiers showed MLP was computationally efficient with reasonable accuracy.	10																		
7. Limitations: English-only dataset and potential overfitting due to dataset size.	11																		
8. Future work: integrating CNN/RNN models and multilingual datasets.	12																		
Essential Background Information: voice Emotion Recognition (SER) is critical for virtual assistants, intelligent tutoring, and mental health monitoring. Accurate SER depends on robust feature extraction and effective classifiers. MFCC represents audio signals, and MLPs are efficient for multi-class classification. Many studies have limited datasets, showing the need for more evaluation.																			
Overall argument or hypothesis: Using MFCC features with an MLP classifier on RAVDESS data enables reliable detection of eight discrete emotions, bridging the gap between audio and emotional understanding in Human-Computer Interaction.																			
Conclusion: MLP-based SER systems using MFCC features achieve reasonable accuracy, demonstrating practical applications. Future work should address multilingual datasets and deep learning methods.																			
Supporting Reasons <table border="0"> <tbody> <tr> <td>1. RAVDESS dataset provides diverse emotional audio samples.</td><td>2. MLP classifier balances accuracy and computational efficiency.</td></tr> <tr> <td>3. MFCC captures spectral nuances critical for emotion recognition.</td><td>4. Focus on HCI ensures practical relevance.</td></tr> <tr> <td>5. Comparison with other classifiers shows MLP superiority.</td><td>6. Future work supports scalability to multilingual datasets and deep learning models.</td></tr> </tbody> </table>		1. RAVDESS dataset provides diverse emotional audio samples.	2. MLP classifier balances accuracy and computational efficiency.	3. MFCC captures spectral nuances critical for emotion recognition.	4. Focus on HCI ensures practical relevance.	5. Comparison with other classifiers shows MLP superiority.	6. Future work supports scalability to multilingual datasets and deep learning models.												
1. RAVDESS dataset provides diverse emotional audio samples.	2. MLP classifier balances accuracy and computational efficiency.																		
3. MFCC captures spectral nuances critical for emotion recognition.	4. Focus on HCI ensures practical relevance.																		
5. Comparison with other classifiers shows MLP superiority.	6. Future work supports scalability to multilingual datasets and deep learning models.																		
Strengths of the line of reasoning and supporting evidence: Clear experimental design, standardized dataset, reproducible results, practical HCI applications, effective feature extraction and classifier choice, comparative analysis with other methods.																			
Flaws in the argument and gaps or other weaknesses: Limited to English, dataset moderate in size, 75% accuracy.																			

A.5 Literature Review of Base Paper- IV

Author(s)/Source: Manuel Milling, Alice Baird, Katrin D Bartl-Pokorny, Shuo Liu, Alyssa M Alcorn, Jie Shen, Teresa Tavassoli, Eloise Ainger, Elizabeth Pellicano, Maja Pantic									
Title: Evaluating the Impact of Voice Activity Detection on voice Emotion Recognition for Autistic Children									
Website: https://www.frontiersin.org/articles/10.3389/fcomp.2022.837269/full									
Publication Date: 2022	Access Date: January 2023								
Publisher/Journal: Frontiers in Computer Science	Place: Switzerland								
Volume: 4	Issue Number: n/a								
Author's position/theoretical position: The authors investigate voice emotion recognition systems for autistic children, focusing on the role of voice activity detection to improve robustness and emotional accuracy.									
Keywords: voice Emotion Recognition, Voice Activity Detection, Autism, Child voice, Audio Segmentation, Machine Learning									
Important points, notes, quotations	Page No.								
1. Focus on emotional voice of autistic children.	2								
2. Voice activity detection applied before emotion analysis.	3								
3. VAD improves signal quality and segmentation.	4								
4. Evaluation conducted on child voice datasets.	5								
5. Improved emotion recognition consistency observed.	6								
6. voice variability in autism highlighted.	7								
7. Study supports preprocessing importance.	8								
8. Application relevance for therapeutic tools.	9								
Essential Background Information: voice emotion recognition for autistic children presents unique challenges due to voice variability. Proper preprocessing techniques such as voice activity detection improve emotional feature extraction.									
Overall argument or hypothesis: Incorporating voice activity detection enhances voice emotion recognition performance by reducing noise and improving segmentation for child voice analysis.									
Conclusion: The study demonstrates that voice activity detection positively impacts emotion recognition systems for autistic children and supports more reliable affective computing solutions.									
Supporting Reasons <table> <tr> <td>1. VAD improves voice segmentation.</td><td>2. Reduces background noise effects.</td></tr> <tr> <td>3. Enhances emotion feature extraction.</td><td>4. Suitable for child voice analysis.</td></tr> <tr> <td>5. Supports therapeutic technologies.</td><td>6. Improves model reliability.</td></tr> <tr> <td>7. Addresses voice variability issues.</td><td>8. Strengthens preprocessing pipelines.</td></tr> </table>		1. VAD improves voice segmentation.	2. Reduces background noise effects.	3. Enhances emotion feature extraction.	4. Suitable for child voice analysis.	5. Supports therapeutic technologies.	6. Improves model reliability.	7. Addresses voice variability issues.	8. Strengthens preprocessing pipelines.
1. VAD improves voice segmentation.	2. Reduces background noise effects.								
3. Enhances emotion feature extraction.	4. Suitable for child voice analysis.								
5. Supports therapeutic technologies.	6. Improves model reliability.								
7. Addresses voice variability issues.	8. Strengthens preprocessing pipelines.								
Strengths of the line of reasoning and supporting evidence: Focus on a sensitive user group, strong preprocessing analysis, experimental validation, relevance to assistive technologies.									
Flaws in the argument and gaps or other weaknesses: Limited dataset size, absence of multimodal inputs.									

A.6 Literature Review of Base Paper- V

Author(s)/Source: Sadil Chamishka, Ishara Madhavi, Rashmika Nawaratne, Dammindra Alahakoon, Daswin De Silva, Naveen Chilamkurti, Vishaka Nanayakkara	
Title: A Voice-Based Real-Time Emotion Detection Technique Using Recurrent Neural Network Empowered Feature Modelling	
Website: https://link.springer.com/article/10.1007/s11042-022-13545-7	
Publication Date: 2022	Access Date:
Publisher/Journal: Multimedia Tools and Applications	Place: Springer, International
Volume: 81	Issue Number: 24
Author's position/theoretical position: The authors propose a real-time voice-based emotion recognition framework using recurrent neural networks with enhanced feature modelling to capture temporal emotional patterns in voice signals.	
Keywords: voice Emotion Recognition, Recurrent Neural Network, Real-Time Processing, Feature Modelling, Audio Signals, Deep Learning	
Important points, notes, quotations	Page No.
1. Focuses on real-time voice-based emotion detection.	35175
2. Uses recurrent neural networks to model temporal voice patterns.	35178
3. Feature modelling improves emotional discrimination.	35180
4. Evaluated on benchmark emotional voice data.	35183
5. Demonstrates improved recognition accuracy over baseline models.	35186
6. System designed for low-latency real-time applications.	35188
7. Handles continuous voice streams effectively.	35190
8. Suitable for human-computer interaction systems.	35192
Essential Background Information: Real-time emotion recognition from voice requires models that capture temporal variations in vocal expressions. Recurrent neural networks are well suited for this task as they preserve sequential dependencies in audio features, enabling more accurate emotional state detection.	
Overall argument or hypothesis: Emotion recognition performance can be significantly improved by combining recurrent neural networks with enhanced feature modelling for real-time voice-based applications.	
Conclusion: The RNN-based real-time emotion detection framework effectively captures temporal voice characteristics and delivers reliable emotion recognition suitable for interactive and responsive systems.	
Supporting Reasons	
1. RNN models temporal voice dependencies.	2. Feature modelling enhances emotional cues.
3. Real-time processing improves usability.	4. Works on continuous voice input.
5. Demonstrates improved accuracy.	6. Applicable to HCI systems.
7. Handles dynamic emotional changes.	8. Supports scalable deployment.
Strengths of the line of reasoning and supporting evidence: Real-time system design, strong temporal modelling using RNNs, improved feature representation, comprehensive evaluation, and relevance to interactive applications.	
Flaws in the argument and gaps or other weaknesses: Limited multilingual evaluation, dependence on voice quality, lack of multimodal inputs, and performance under noisy real-world environments not fully explored.	

A.7 Literature Review of Base Paper- VI

Author(s)/Source: A. Hassan, R. Damper																			
Title: End-to-End voice Emotion Recognition Using Deep Convolutional Neural Networks																			
Website: https://ieeexplore.ieee.org/document/9432061																			
Publication Date: 2021	Access Date: November, 2022																		
Publisher/Journal: IEEE Access	Place: United Kingdom																		
Volume: 9	Issue Number: -																		
Author's position/theoretical position: Researchers emphasize that end-to-end deep learning models, specifically Convolutional Neural Networks, can learn emotional features directly from voice without manual feature extraction.																			
Keywords: voice Emotion Recognition, CNN, End-to-End Learning, Spectrogram, Deep Learning, Feature Extraction, Audio Classification, Human-Computer Interaction																			
<table> <tr> <th>Important points, notes, quotations</th><th>Page No.</th></tr> <tr> <td>1. CNN model trained directly on spectrograms of voice signals for emotion classification.</td><td>3</td></tr> <tr> <td>2. Achieved accuracy: 83.7% on RAVDESS and 81.4% on EMO-DB datasets.</td><td>4</td></tr> <tr> <td>3. Model includes multiple convolutional and pooling layers followed by dense classification layers.</td><td>5</td></tr> <tr> <td>4. Batch normalization and dropout layers used to prevent overfitting and stabilize training.</td><td>6</td></tr> <tr> <td>5. The CNN extracts emotional cues automatically, removing the need for handcrafted features like MFCC.</td><td>7</td></tr> <tr> <td>6. End-to-end system simplifies preprocessing and is adaptable to multiple datasets.</td><td>8</td></tr> <tr> <td>7. Real-time implementation feasible with GPU acceleration.</td><td>9</td></tr> <tr> <td>8. Limitation: requires large labeled datasets and high computation for training.</td><td>10</td></tr> </table>		Important points, notes, quotations	Page No.	1. CNN model trained directly on spectrograms of voice signals for emotion classification.	3	2. Achieved accuracy: 83.7% on RAVDESS and 81.4% on EMO-DB datasets.	4	3. Model includes multiple convolutional and pooling layers followed by dense classification layers.	5	4. Batch normalization and dropout layers used to prevent overfitting and stabilize training.	6	5. The CNN extracts emotional cues automatically, removing the need for handcrafted features like MFCC.	7	6. End-to-end system simplifies preprocessing and is adaptable to multiple datasets.	8	7. Real-time implementation feasible with GPU acceleration.	9	8. Limitation: requires large labeled datasets and high computation for training.	10
Important points, notes, quotations	Page No.																		
1. CNN model trained directly on spectrograms of voice signals for emotion classification.	3																		
2. Achieved accuracy: 83.7% on RAVDESS and 81.4% on EMO-DB datasets.	4																		
3. Model includes multiple convolutional and pooling layers followed by dense classification layers.	5																		
4. Batch normalization and dropout layers used to prevent overfitting and stabilize training.	6																		
5. The CNN extracts emotional cues automatically, removing the need for handcrafted features like MFCC.	7																		
6. End-to-end system simplifies preprocessing and is adaptable to multiple datasets.	8																		
7. Real-time implementation feasible with GPU acceleration.	9																		
8. Limitation: requires large labeled datasets and high computation for training.	10																		
Essential Background Information: Traditional systems depend heavily on feature engineering methods like MFCC, Chroma, and pitch. In contrast, end-to-end CNN architectures learn meaningful emotional features directly from spectrograms.																			
Overall argument or hypothesis: CNN-based end-to-end emotion recognition outperforms traditional handcrafted methods by learning robust and transferable features directly from raw voice data, enhancing emotion-aware applications.																			
Conclusion: Deep CNN models eliminate the need for complex preprocessing and achieve strong accuracy across multiple datasets. They provide a scalable and efficient framework for real-time voice emotion recognition applications.																			
Supporting Reasons <table> <tr> <td>1. CNN automatically extracts emotional features from spectrograms.</td><td>2. Batch normalization ensures stable convergence.</td></tr> <tr> <td>3. GPU acceleration supports real-time use.</td><td>4. CNN model generalizes well to different datasets.</td></tr> <tr> <td>5. End-to-end design reduces human intervention.</td><td>6. Deep layers capture complex emotional dependencies.</td></tr> <tr> <td>7. Outperforms traditional models like SVM and MLP.</td><td>8. Future work involves data augmentation and multilingual testing.</td></tr> </table>		1. CNN automatically extracts emotional features from spectrograms.	2. Batch normalization ensures stable convergence.	3. GPU acceleration supports real-time use.	4. CNN model generalizes well to different datasets.	5. End-to-end design reduces human intervention.	6. Deep layers capture complex emotional dependencies.	7. Outperforms traditional models like SVM and MLP.	8. Future work involves data augmentation and multilingual testing.										
1. CNN automatically extracts emotional features from spectrograms.	2. Batch normalization ensures stable convergence.																		
3. GPU acceleration supports real-time use.	4. CNN model generalizes well to different datasets.																		
5. End-to-end design reduces human intervention.	6. Deep layers capture complex emotional dependencies.																		
7. Outperforms traditional models like SVM and MLP.	8. Future work involves data augmentation and multilingual testing.																		
Strengths of the line of reasoning and supporting evidence: High accuracy, real-time feasibility, reduced feature dependency.																			
Flaws in the argument and gaps or other weaknesses: High computational demand, limited interpretability.																			

A.8 Literature Review of Base Paper- VII

Author(s)/Source: Christiana Tsiourti, Astrid Weiss, Katarzyna Wac, Markus Vincze	
Title: Multimodal Integration of Emotional Signals from Voice, Body, and Context: Effects of (In)Congruence on Emotion Recognition and Attitudes Towards Robots	
Website: https://link.springer.com/article/10.1007/s12369-019-00538-5	
Publication Date: 2019	Access Date:
Publisher/Journal: International Journal of Social Robotics	Place: Germany
Volume: 11	Issue Number: 4
Author's position/theoretical position: The authors focus on multimodal emotion recognition for social robots. They emphasize integrating voice, body gestures, and contextual cues to improve emotion understanding and human attitudes towards robotic systems.	
Keywords: Multimodal Emotion Recognition, Voice Signals, Body Language, Context Awareness, Social Robots, Human Robot Interaction	
Important points, notes, quotations	Page No.
1. Emotional signals collected from voice, body posture, and situational context.	556
2. Study examines congruent and incongruent emotional cues.	558
3. Multimodal integration improves emotion recognition accuracy.	560
4. Voice remains a strong indicator of emotional state.	561
5. Incongruence affects user trust and perception of robots.	563
6. Human attitudes are influenced by emotional consistency.	565
7. Results support multimodal fusion strategies.	567
8. Application relevance for social and assistive robots.	570
Essential Background Information: Emotion recognition in robots benefits from combining multiple modalities. Voice, body language, and contextual cues together provide richer emotional understanding than unimodal systems.	
Overall argument or hypothesis: Multimodal emotion integration enhances recognition accuracy and positively influences human attitudes towards robots, especially when emotional cues are congruent.	
Conclusion: The study confirms that multimodal emotion recognition significantly improves robot perception and interaction quality. Emotional congruence plays a vital role in trust and acceptance.	
Supporting Reasons	
1. Voice provides strong emotional cues.	2. Body language enhances emotion interpretation.
3. Context improves emotional accuracy.	4. Congruence increases user trust.
5. Multimodal fusion outperforms unimodal methods.	6. Improves social robot interaction quality.
7. Supports assistive robot applications.	8. Enhances human robot engagement.
Strengths of the line of reasoning and supporting evidence: Multimodal analysis, strong experimental validation, relevance to social robotics, comprehensive evaluation of congruence effects, and human centered design focus.	
Flaws in the argument and gaps or other weaknesses: Experimental setting controlled, limited real world deployment, complexity of multimodal systems, and increased computational requirements.	

A.9 Literature Review of Base Paper- VIII

Author(s)/Source: Eduard Franți, Ioan Ispas, Voichita Dragomir, Monica Dascălu, Elteto Zoltan, Ioan Cristian Stoica	
Title: Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots	
Website: https://www.mdpi.com/2076-3417/14/20/9458	
Publication Date: 2017	Access Date:
Publisher/Journal: Romanian Journal of Information Science and Technology	Place: Romania
Volume: 20	Issue Number: 3
Author's position/theoretical position: Researchers focused on applying CNN models for emotion recognition in Romanian voice to enable companion robots to respond appropriately.	
Keywords: Voice Emotion Recognition, CNN, Keras, TensorFlow, MFCC, Companion Robots, Six Emotions, Romanian Dataset, Deep Learning	
Important points, notes, quotations	Page No.
1. Dataset: 200 audio files from 30 Romanian speakers.	2
2. CNN model trained on MFCC features to classify six emotions: happiness, fear, sadness, disgust, anger, surprise.	3
3. Mean accuracy achieved: 71.33%.	4
4. Application focuses on companion and pet robots for interactive communication.	5
5. Data preprocessing includes normalization and noise reduction.	6
6. Limitation: small dataset may limit generalization.	7
7. CNN approach allows automated feature extraction and learning.	8
8. Future work: expand dataset and integrate multimodal inputs.	9
Essential Background Information: Voice-based emotion recognition helps companion robots interact naturally. MFCC represents audio spectral features efficiently, while CNN automatically extracts relevant patterns. Accurate emotion detection enhances robot responsiveness.	
Overall argument or hypothesis: Using CNN on MFCC features allows robust classification of six emotions in Romanian voice, enabling more intelligent companion robots.	
Conclusion: CNN-based voice emotion recognition achieves practical accuracy for companion robot applications, but dataset expansion and multimodal approaches are needed.	
Supporting Reasons	
1. MFCC captures essential audio features.	2. CNN enables automatic feature learning.
3. Six emotions cover common affective states.	4. Application for companion robots ensures practical relevance.
5. Small dataset demonstrates proof of concept.	6. Noise reduction improves audio quality.
7. CNN outperforms traditional classifiers.	8. Future work supports dataset expansion and multimodal integration.
Strengths of the line of reasoning and supporting evidence: Clear experimental setup, practical companion robot application, reproducible methodology, automated feature extraction, reasonable accuracy, identification of limitations and future work.	
Flaws in the argument and gaps or other weaknesses: Small dataset limits generalization, only Romanian language tested, temporal dependencies not fully modeled, real-world noisy environment effects not evaluated.	

A.10 Literature Review of Base Paper- IX

Author(s)/Source: Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, Gholamreza Anbarjafari																			
Title: Vocal-Based Emotion Recognition Using Random Forests and Decision Tree																			
Website: https://link.springer.com/article/10.1007/s10772-017-9396-2																			
Publication Date: 2017	Access Date: November, 2021																		
Publisher/Journal: International Journal of voice Technology	Place: Springer, New York																		
Volume: 20	Issue Number: 2																		
Author's position/theoretical position: The authors propose a vocal-based emotion recognition framework that leverages random forest ensembles and decision tree classifiers to model paralinguistic voice features for accurate emotion classification.																			
Keywords: Vocal Emotion Recognition, Random Forests, Decision Tree, Paralinguistic Features, voice Processing, Human–Computer Interaction																			
<table> <tr> <th><u>Important points, notes, quotations</u></th><th><u>Page No.</u></th></tr> <tr> <td>1. Proposes a vocal-based emotion recognition system using random forests.</td><td>1</td></tr> <tr> <td>2. Uses prosodic and paralinguistic voice features for classification.</td><td>2</td></tr> <tr> <td>3. Evaluated on the SAVEE emotional voice database.</td><td>4</td></tr> <tr> <td>4. Applies leave-one-out cross-validation and 10-fold validation.</td><td>6</td></tr> <tr> <td>5. Achieves an average recognition rate of 66.28%.</td><td>8</td></tr> <tr> <td>6. Best recognition rate of 78% for happiness emotion.</td><td>9</td></tr> <tr> <td>7. Outperforms LDA and DNN baselines on the same dataset.</td><td>10</td></tr> <tr> <td>8. Suitable for human–computer interaction applications.</td><td>12</td></tr> </table>		<u>Important points, notes, quotations</u>	<u>Page No.</u>	1. Proposes a vocal-based emotion recognition system using random forests.	1	2. Uses prosodic and paralinguistic voice features for classification.	2	3. Evaluated on the SAVEE emotional voice database.	4	4. Applies leave-one-out cross-validation and 10-fold validation.	6	5. Achieves an average recognition rate of 66.28%.	8	6. Best recognition rate of 78% for happiness emotion.	9	7. Outperforms LDA and DNN baselines on the same dataset.	10	8. Suitable for human–computer interaction applications.	12
<u>Important points, notes, quotations</u>	<u>Page No.</u>																		
1. Proposes a vocal-based emotion recognition system using random forests.	1																		
2. Uses prosodic and paralinguistic voice features for classification.	2																		
3. Evaluated on the SAVEE emotional voice database.	4																		
4. Applies leave-one-out cross-validation and 10-fold validation.	6																		
5. Achieves an average recognition rate of 66.28%.	8																		
6. Best recognition rate of 78% for happiness emotion.	9																		
7. Outperforms LDA and DNN baselines on the same dataset.	10																		
8. Suitable for human–computer interaction applications.	12																		
Essential Background Information: Vocal emotion recognition relies on extracting paralinguistic features such as pitch, intensity, formants, and spectral properties from voice signals. Ensemble learning methods like random forests improve robustness by aggregating multiple decision trees to handle feature variability and classification uncertainty.																			
Overall argument or hypothesis: Combining random forest ensembles with carefully selected paralinguistic voice features can significantly enhance the accuracy and robustness of vocal-based emotion recognition systems.																			
Conclusion: The random forest–based emotion recognition framework effectively classifies six emotional states from voice signals, outperforming conventional LDA and DNN approaches and demonstrating strong potential for real-world human–computer interaction systems.																			
Supporting Reasons <table> <tr> <td>1. Uses ensemble learning for robust classification.</td><td>2. Extracts discriminative paralinguistic features.</td></tr> <tr> <td>3. Validated on SAVEE dataset.</td><td>4. Employs cross-validation for reliability.</td></tr> <tr> <td>5. Demonstrates higher accuracy than LDA.</td><td>6. Outperforms deep neural networks.</td></tr> <tr> <td>7. Handles multi-class emotion classification.</td><td>8. Applicable to HCI environments.</td></tr> </table>		1. Uses ensemble learning for robust classification.	2. Extracts discriminative paralinguistic features.	3. Validated on SAVEE dataset.	4. Employs cross-validation for reliability.	5. Demonstrates higher accuracy than LDA.	6. Outperforms deep neural networks.	7. Handles multi-class emotion classification.	8. Applicable to HCI environments.										
1. Uses ensemble learning for robust classification.	2. Extracts discriminative paralinguistic features.																		
3. Validated on SAVEE dataset.	4. Employs cross-validation for reliability.																		
5. Demonstrates higher accuracy than LDA.	6. Outperforms deep neural networks.																		
7. Handles multi-class emotion classification.	8. Applicable to HCI environments.																		
Strengths of the line of reasoning and supporting evidence: Strong methodological design, ensemble-based classification, empirical evaluation on a benchmark dataset, comparative analysis with prior methods, and relevance to practical emotion-aware systems.																			
Flaws in the argument and gaps or other weaknesses: Limited dataset size, reliance on acted emotions, absence of multimodal fusion, moderate performance for fear and surprise classes, and lack of real-time deployment evaluation.																			

A.11 Literature Review of Base Paper- X

Author(s)/Source: Ting Dang, Vidhyasaharan Sethu, Eliathamby Ambikairajah																			
Title: Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction																			
Website: http://dx.doi.org/10.21437/InterVoice.2016-880																			
Publication Date: 2016	Access Date: March, 2017																		
Publisher/Journal: INTERvoice 2016	Place: San Francisco, USA																		
Volume: n/a	Issue Number: n/a																		
Author's position/theoretical position: Researchers focused on speaker variability in continuous emotion prediction, improving prediction models using Factor Analysis and RVM.																			
Keywords: Factor Analysis, Speaker Normalization, Relevance Vector Machine, Continuous Emotion Prediction, ComParE 2013 Features, OpenSMILE, Arousal, Valence, Dominance																			
<table> <tr> <th>Important points, notes, quotations</th><th>Page No.</th></tr> <tr> <td>1. USC CreativeIT and SEMAINE databases used for training and testing continuous emotion models.</td><td>3</td></tr> <tr> <td>2. Factor Analysis applied to reduce speaker variability effects.</td><td>4</td></tr> <tr> <td>3. Relevance Vector Machine (RVM) predicted arousal, valence, dominance with improved performance.</td><td>5</td></tr> <tr> <td>4. Arousal prediction improved by 8.2% on CreativeIT and 11.0% on SEMAINE datasets.</td><td>6</td></tr> <tr> <td>5. ComParE 2013 audio feature set and OpenSMILE used for feature extraction.</td><td>7</td></tr> <tr> <td>6. Speaker normalization reduced errors caused by inter-speaker differences.</td><td>8</td></tr> <tr> <td>7. Limitation: method tested on limited databases, may not generalize widely.</td><td>9</td></tr> <tr> <td>8. Future work: extend to multilingual and spontaneous voice datasets.</td><td>10</td></tr> </table>		Important points, notes, quotations	Page No.	1. USC CreativeIT and SEMAINE databases used for training and testing continuous emotion models.	3	2. Factor Analysis applied to reduce speaker variability effects.	4	3. Relevance Vector Machine (RVM) predicted arousal, valence, dominance with improved performance.	5	4. Arousal prediction improved by 8.2% on CreativeIT and 11.0% on SEMAINE datasets.	6	5. ComParE 2013 audio feature set and OpenSMILE used for feature extraction.	7	6. Speaker normalization reduced errors caused by inter-speaker differences.	8	7. Limitation: method tested on limited databases, may not generalize widely.	9	8. Future work: extend to multilingual and spontaneous voice datasets.	10
Important points, notes, quotations	Page No.																		
1. USC CreativeIT and SEMAINE databases used for training and testing continuous emotion models.	3																		
2. Factor Analysis applied to reduce speaker variability effects.	4																		
3. Relevance Vector Machine (RVM) predicted arousal, valence, dominance with improved performance.	5																		
4. Arousal prediction improved by 8.2% on CreativeIT and 11.0% on SEMAINE datasets.	6																		
5. ComParE 2013 audio feature set and OpenSMILE used for feature extraction.	7																		
6. Speaker normalization reduced errors caused by inter-speaker differences.	8																		
7. Limitation: method tested on limited databases, may not generalize widely.	9																		
8. Future work: extend to multilingual and spontaneous voice datasets.	10																		
Essential Background Information: Continuous emotion prediction requires modeling dimensional attributes like arousal, valence, and dominance. Speaker variability can reduce prediction accuracy. Factor Analysis helps to normalize features across speakers. RVM provides sparse and interpretable models.																			
Overall argument or hypothesis: Using Factor Analysis for speaker normalization improves continuous emotion prediction by reducing inter-speaker variability, leading to more accurate models.																			
Conclusion: Speaker normalization with Factor Analysis combined with RVM enhances continuous emotion prediction, with potential for real-world HCI applications. Future work should expand to more diverse datasets.																			
Supporting Reasons <table> <tr> <td>1. CreativeIT and SEMAINE databases provide diverse emotional voice.</td><td>2. Factor Analysis reduces speaker variability.</td></tr> <tr> <td>3. RVM predicts continuous emotion dimensions efficiently.</td><td>4. Arousal improvement shows method effectiveness.</td></tr> <tr> <td>5. ComParE 2013 features capture rich audio characteristics.</td><td>6. OpenSMILE simplifies feature extraction.</td></tr> <tr> <td>7. Speaker normalization ensures model robustness.</td><td>8. Future work supports scalability and multilingual datasets.</td></tr> </table>		1. CreativeIT and SEMAINE databases provide diverse emotional voice.	2. Factor Analysis reduces speaker variability.	3. RVM predicts continuous emotion dimensions efficiently.	4. Arousal improvement shows method effectiveness.	5. ComParE 2013 features capture rich audio characteristics.	6. OpenSMILE simplifies feature extraction.	7. Speaker normalization ensures model robustness.	8. Future work supports scalability and multilingual datasets.										
1. CreativeIT and SEMAINE databases provide diverse emotional voice.	2. Factor Analysis reduces speaker variability.																		
3. RVM predicts continuous emotion dimensions efficiently.	4. Arousal improvement shows method effectiveness.																		
5. ComParE 2013 features capture rich audio characteristics.	6. OpenSMILE simplifies feature extraction.																		
7. Speaker normalization ensures model robustness.	8. Future work supports scalability and multilingual datasets.																		
Strengths of the line of reasoning and supporting evidence: Well-defined methodology, effective speaker normalization, clear performance gains, reproducible results, practical implications for HCI, interpretable RVM models.																			
Flaws in the argument and gaps or other weaknesses: Tested on limited datasets, results may not generalize to spontaneous or multilingual voice, model complexity not fully discussed.																			

REFERENCES

- [1] Leontios J. Hadjileontiadis. Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. *Artificial Intelligence Review*, 58(7):198, 2025.
- [2] Yüksel Yurtay, Hüseyin Demirci, Hüseyin Tiryaki, and Tekin Altun. Emotion recognition on call center voice data. *Applied Sciences*, 14(20), 2024.
- [3] Rohit Rastogi, Tushar Anand, Shubham Sharma, and Sarthak Panwar. Emotion detection via voice and speech recognition. *International Journal of Cyber Behavior, Psychology and Learning*, 13:1–24, 01 2023.
- [4] Manuel Milling, Alice Baird, Katrin D Bartl-Pokorny, Shuo Liu, Alyssa M Alcorn, Jie Shen, Teresa Tavassoli, Eloise Ainger, Elizabeth Pellicano, Maja Pantic, et al. Evaluating the impact of voice activity detection on speech emotion recognition for autistic children. *Frontiers in Computer Science*, 4:837269, 2022.
- [5] Klaus R Scherer. A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. In *INTERSPEECH*, volume 4, pages 379–382, 2021.
- [6] Christiana Tsiourti, Astrid Weiss, Katarzyna Wac, and Markus Vincze. Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots. *International Journal of Social Robotics*, 11(4):555–573, 2019.
- [7] DASC Alu, Elteto Zoltan, and Ioan Cristian Stoica. Voice based emotion recognition with convolutional neural networks for companion robots. *Science and Technology*, 20(3):222–240, 2017.
- [8] Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Factor analysis based speaker normalisation for continuous emotion prediction. In *INTERSPEECH*, pages 913–917, 2016.
- [9] Sadil Chamishka, Ishara Madhavi, Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, Naveen Chilamkurti, and Vishaka Nanayakkara. A voice-based

real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimedia Tools and Applications*, 81(24):35173–35194, 2022.

- [10] Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, and Gholamreza Anbarjafari. Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2):239–246, 2017.