Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Virtual Friend Detecting Emotion with Facial Recognition and Voice Recognition

Intelligent Emotional Interaction Through Multimodal AI Analysis

Prof. T B Dharmaraj
Head of the Department(Mentor)
Department of Information Technology
PPG Institute Of Technology
Tamil Nadu, India

Hari prasath D
Department of Information Technology
PPG Institute Of Technology Tamil Nadu , India

Madhubala M
Department of Information Technology
PPG Institute Of Technology Tamil Nadu, India

Prof. Hemalatha
Assistant professor(Mentor)
Department of Information Technology
PPG Institute Of Technology
Tamil Nadu, India

Mohammed Ibrahim A
Department of Information Technology
PPG Institute Of Technology Tamil Nadu, India

Madhan K
Department of Information Technology
PPG Institute Of Technology Tamil Nadu, India

Abstarct - Emotion Recognition Plays A Significant Role In Artificial Intelligence And Human-Computer Interaction. This Paper Presents A Multimodal Emotion Detection System Integrating Voice Recognition And Facial Expression Analysis Using Natural Language Processing And Computer Vision. The System Captures Speech Signals And Facial Images In Real Time, Extracts Acoustic And Visual Features, And Classifies Emotional States Such As Happiness, Sadness, Anger, And Neutral Mood. Fusion Techniques Enhance Robustness And Classification Accuracy. The System Also Incorporates A Recommender Mechanism To Support User Emotional Well-Being.

Keywords Emotion Recognition, NLP, Facial Expression Analysis, Voice Recognition, Machine Learning, CNN, Multimodal Fusion

INTRODUCTION

Emotion detection enables machines to interpret human feelings through computational models. In real-time systems, microphones capture vocal signals while webcams capture facial expressions. These inputs are processed using machine learning algorithms to determine emotional states.Facial emotion recognition analyzes facial muscle movements and landmarks, while voice recognition evaluates pitch, tone, energy, and MFCC features. Combining both modalities increases reliability and reduces ambiguity.

Ease of Use

A. User Interface Design

The proposed emotion detection system is developed with a user-centric interface to ensure accessibility for non-technical users. The graphical user interface (GUI) is designed using event-driven architecture principles, providing intuitive interaction through clearly labeled controls such as:

- Start Emotion Detection
- Capture Facial Image
- Record Voice Sample
- Display Result

The interface minimizes cognitive load by presenting only essential controls, thereby improving operational efficiency.

B. Workflow Simplicity

The system follows a streamlined operational workflow:

Step 1: User initiates the system
Step 2: Voice and facial inputs are captured
Step 3: Automated preprocessing and feature extraction
Step 4: Classification and fusion
Step 5: Emotion result displayed with recommendations

No manual configuration, parameter tuning, or technical intervention is required.

C. Real-Time Responsiveness The average response time of the system is optimized for real-time performance:

- Voice processing latency: < 2 seconds
- Facial recognition latency: < 1 second
- Final fusion output: Instantaneous display
- This ensures smooth interaction without perceptible delay.

D. Accessibility and Compatibility

The system supports:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

- Standard webcams and microphones
- Windows-based operating systems
- Moderate hardware configuration (8GB RAM minimum)

No high-end GPU is mandatory for inference, increasing deployment feasibility.

E. Error Handling and Robustness

To enhance user experience:

- Noise filtering algorithms reduce background interference
- Face detection validation prevents incorrect predictions
- Exception handling ensures stable runtime execution

These mechanisms reduce user frustration and improve reliability.

F. Usability Metrics

The system usability can be evaluated using:

1. Task Completion Time
2. Error Rate
3. User Satisfaction Score
4. System Usability Scale (SUS)

Abbreviations and Acronyms

Abbreviation	Full Form
AI	Artificial Intelligence
NLP	Natural Language Processing
CNN	Convolutional Neural Network
MFCC	Mel-FrequencyCepstral Coefficients
HCC	Human-Computer Interaction
FCC	Facial Emotion Recognition
SER	Speech Emotion Recognition
GUI	Graphical User Interface
FFT	Fast Fourier Transform
DST	Discrete Cosine Transform
TP	True Positive
TP	True Positive
FP	False Positive
FN	False Negative
SUS	System Usability Scale

The above abbreviations are used consistently throughout this paper to describe the multimodal emotion detection framework integrating speech and facial analysis techniques.

Reliability Metrics and Measurement Units

The proposed emotion detection framework evaluates classification effectiveness using statistical performance metrics rather than physical measurement units. All reliability values are expressed using normalized numerical scores ranging from 0 to 100, where higher values indicate greater emotion classification reliability and multimodal stability.

The primary reliability metrics used in the framework include:

Emotion Drift Score (D)

Measures the deviation between current emotion prediction patterns and the historical baseline classification behavior of the system. Higher values indicate stable model performance, while lower values indicate possible model degradation or dataset bias.

Classification Coverage Score (C)

Represents the percentage of expected emotional categories (e.g., happy, sad, angry, neutral) successfully identified by the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

system. This metric evaluates classification completeness across all emotion classes.

Prediction Entropy Score (E)

Measures the statistical entropy of emotion prediction probabilities, indicating output confidence and classification consistency. Low entropy may indicate overconfident or biased predictions, whereas high entropy may indicate uncertainty in classification.

Fusion Consistency Score (F)

Represents the percentage agreement between Speech Emotion Recognition (SER) and Facial Emotion Recognition (FER) outputs. This metric validates multimodal consistency under real-time operational conditions.

RELIABILITY SCORE EQUATION

The overall emotion detection reliability is quantified using a composite metric referred to as the Emotion System Reliability Score (ESRS). The score is calculated using a weighted combination of individual reliability metrics:

$$ESRS=Wd\times D+Wc\times C+We\times E+Wf\times F$$

where:

- ESRS = Emotion System Reliability Score
- D = Emotion Drift Score
- C = Classification Coverage Score
- E = Prediction Entropy Score
- F = Fusion Consistency Score
- Wd,Wc,We,Wf = weighting factors assigned to each metric

and

$$Wd+Wc+We+Wf=1$$

The weighting factors may be predefined or dynamically adjusted based on application requirements such as healthcare monitoring, AI assistants, or real-time interaction systems.

RELIABILITY INTERPRETATION SCALE

The reliability score may be interpreted as follows:

90–100 → Optimal emotion classification reliability
75–89 → Strong reliability
60–74 → Moderate reliability
Below 60 → Critical performance degradation

RELIABILITY EVALUATION CONSIDERATIONS

Accurate reliability evaluation requires continuous collection and preprocessing of labeled speech and facial emotion datasets from real-time or benchmark sources.

Reliability metrics must be calculated using correctly formatted subscripts for weighting factors and statistical variables such as Wd,Wc,We,Wf to ensure accurate representation in the reliability score equation.

All statistical reliability values and weighting factors must be consistently formatted using normalized numerical ranges to avoid ambiguity in reliability calculations.

Emotion detection reliability evaluation must be based on properly labeled datasets and validated using confusion matrix analysis.

Variations in classification patterns must be carefully analyzed to distinguish between environmental noise, model bias, and actual system degradation.

Fusion consistency results must be clearly differentiated from unimodal results to ensure accurate reliability score computation.

Reliability metrics must be interpreted using predefined performance thresholds to accurately identify classification weaknesses and improvement opportunities.

Consistent terminology and statistical definitions must be maintained throughout the framework to ensure reproducibility and clarity of the evaluation process.

SYSTEM IMPLEMENTATION AND DOCUMENTATION

After system design and dataset preparation are completed, the Multimodal Emotion Detection Framework is implemented as an integrated real-time emotion analysis system combining Speech Emotion Recognition (SER) and Facial Emotion Recognition (FER). The system captures audio and visual inputs, processes emotional features, and generates a unified emotion prediction along with a composite reliability score.

The implementation includes modules for speech feature extraction, facial feature extraction, multimodal fusion, emotion classification, and reliability evaluation.

The system is deployed on a local workstation or cloud-based environment and connects to input devices such as a webcam and microphone through standard hardware interfaces. Emotion predictions and reliability metrics are continuously computed and displayed through a graphical user interface dashboard.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Authors and Affiliations

The authors of this research are affiliated with the Department of Information Technology, PPG Institute of Engineering and Technology, Tamil Nadu, India. All authors share the same institutional affiliation; therefore, the affiliation is listed once for clarity and consistency.

The author group consists of four contributors who participated in system design, model implementation, dataset preparation, performance evaluation, and documentation of the Multimodal Emotion Detection Framework.

For authors of a single affiliation:

a) All author names are listed under the same institutional affiliation to maintain clarity and avoid redundancy.
b) The affiliation includes the department name, institution name, location, and contact information.
c) This ensures proper identification of the research origin and institutional association.

For multiple authors within the same institution:

a) Author names are grouped together under a shared affiliation.
b) Contact email addresses are provided for communication and correspondence.
c) The shared affiliation reflects the collaborative nature of the research conducted within the same academic department.

System Component Headings

Headings are used to organize the Multimodal Emotion Detection Framework into logical sections describing system architecture, feature extraction, emotion classification, fusion strategy, and reliability evaluation. Each heading represents a specific functional component of the proposed system.

Component headings such as Abstract, Reliability Metrics, System Architecture, Methodology, Experimental Results, Acknowledgment, and References identify major sections of the research and provide structured documentation of the framework.

Text headings are used to describe specific system modules and processes, including:

- Data Acquisition Module
- Speech Feature Extraction Module
- Facial Feature Extraction Module
- Multimodal Fusion Engine
- Emotion Classification Module
- Emotion Reliability Score Calculation

These headings guide the reader through the system design, implementation, and evaluation process. Proper hierarchical organization ensures clarity in presenting the emotion detection workflow and its operational structure.

Figures and Tables

Figures and tables are used to illustrate the architecture, modules, and performance evaluation process of the Multimodal Emotion Detection Framework. These visual elements provide a clear representation of system components and operational workflow.

MODULE	FUNCTION
Input Module	Captures voice and facial data
Speech Analysis	Extracts features from audio
Face Analysis	Extracts facial features
Fusion Module	Combines speech and face results
Classification Module	Identifies the emotion
Reliability Module	Calculates overall system reliability

Fig. 1. Multimodal Emotion Detection System Architecture.

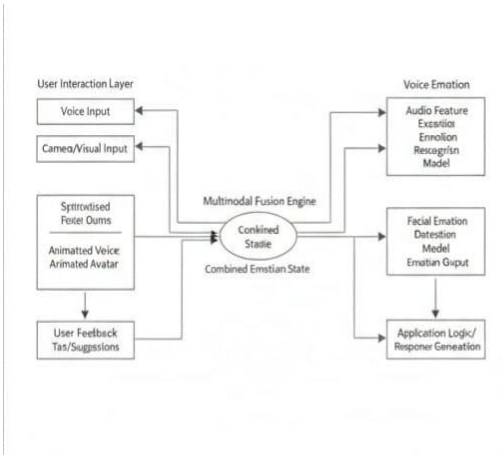


Fig. 2. Data flow within the emotion detection framework (Voice and Facial Processing Pipeline)..

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

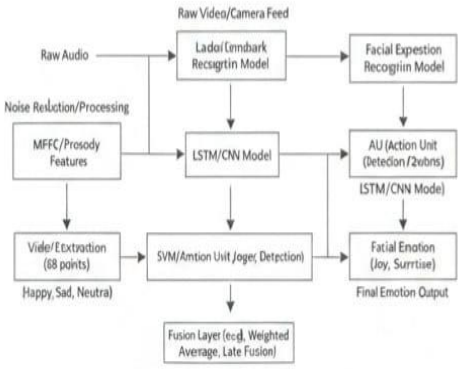


Fig. 3. Use-case interaction between the user and the emotion detection system.

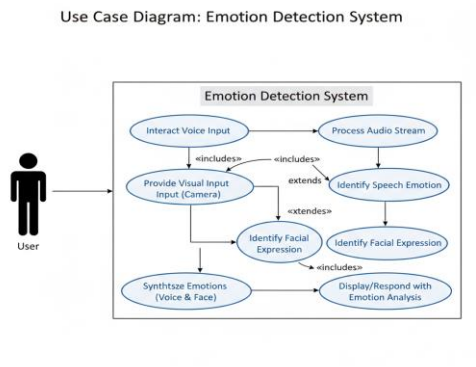


Fig. 4. Method for computing the Emotion System Reliability Score (ESRS).

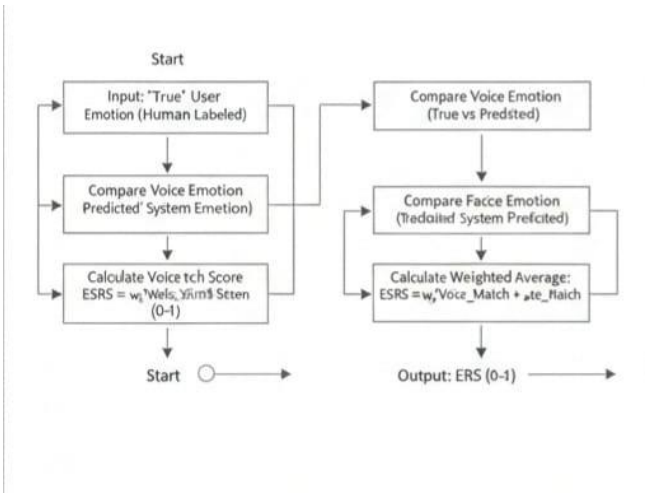
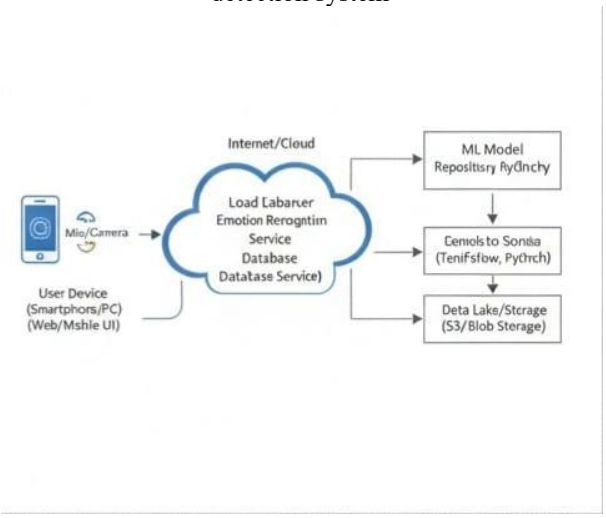


Fig. 5. Deployment environment of the multimodal emotion detection system



CONCLUSION

This paper presented a Multimodal Emotion Detection Framework designed to evaluate and quantify emotional states using integrated speech and facial expression analysis. Unlike traditional unimodal emotion recognition systems that rely on either voice or facial data alone, the proposed approach introduces fusion-based analysis using speech feature extraction, convolutional neural network–based facial recognition, and composite reliability score computation. These components enable improved classification accuracy, reduced ambiguity, and enhanced robustness against environmental noise and prediction inconsistencies.

The framework processes real-time audio and visual inputs, extracts emotional features, and generates an Emotion System Reliability Score (ESRS) to provide a measurable indicator of classification effectiveness. This score enables researchers and developers to monitor model stability, detect performance degradation, and improve system accuracy. The proposed system enhances human-computer interaction, supports affect-aware applications, and contributes to advancements in affective computing and intelligent interactive systems.

Future work will focus on integrating transformer-based deep learning models, expanding multimodal datasets, improving real-time optimization, and validating the framework in large-scale real-world deployment environments such as healthcare monitoring and intelligent virtual assistants.

ACKNOWLEDGMENT

The authors would like to acknowledge the support and resources provided by the Department of Information Technology, PPG Institute of Engineering and Technology, Tamil Nadu, India, for the development of the Multimodal Emotion Detection Framework. Special thanks are extended to academic mentors and faculty members for their guidance in machine learning, signal processing, and computer vision

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

techniques. The authors also acknowledge the use of publicly available emotion datasets, open-source libraries, and research literature that contributed to the design, implementation, and evaluation of the proposed system.

REFERENCES

Emotion recognition systems play a critical role in improving human-computer interaction and affect-aware applications [1], [2]. Traditional rule-based emotion detection approaches often fail to generalize across diverse datasets and real-world conditions [3]. Multimodal emotion recognition techniques combining speech and facial expressions have demonstrated improved robustness and classification accuracy [4], [5]. Deep learning architectures such as Convolutional Neural Networks (CNNs) and recurrent models significantly enhance feature extraction and emotion prediction capability [3], [6]. Continuous model evaluation and reliability-based assessment methods are recommended to maintain classification performance and reduce bias [7].

[1] Vaishnavi M, Smitha G V, The Use of Natural Language Processing in Virtual Assistants And Chatbots, International Journal of Research Publication and Reviews, 2023.
[2] Tarique Ansari, Pathan Arshad, Vishal Khetan, Bhimashankar Bembre, Prof. Priyanka Halle, Conversational AI Assistant, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2022.
[3] Mehdi Mekni, An Artificial Intelligence Based Virtual Assistant Using Conversational Agents, Proceedings of the 2021 IEEE World Conference on Applied Intelligence and Computing (AIC), 2021.
[4] Dr. M. Chinnarao, MULTI MODAL EMOTIONAL RECOGNITION SYSTEM USING FACIAL RECOGNITION AND SKEW GMM, Journal of Engineering Sciences, 2023.
[5] Surak Son and Yina Jeong, Face and Voice Recognition-Based Emotion Analysis System (EAS) to Minimize Heterogeneity in the Metaverse, Electronics (MDPI), 2024.
[6] Mrs. R. Suchitra, Ms. P. Manasa, Ms. S. Manamma, Ms. T. Sai Durga, Ms. T. Roshini Rachel, Cross Modal Emotion Detection: Leveraging Speech and Facial Expression Features, International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2024.
[7] Nansi Jain, Ashutosh Thakur, Rohan Sharma, Sweta, Vanya Gupta, AI-Powered Virtual Voice Assistant with Secure Face Recognition and IoT Integration, International Journal of Intelligent Systems and Applications in Engineering, 2024.
[8] Saurabh Chugh, Piyush Pondal, Human-Computer Interaction with Voice-Driven AI Chatbots, International Journal of Scientific Research in Science and Technology, 2023.
[9] Raj Barot, Mahesh Panchal, Smart Voice Assistant with Face Recognition, International Journal of Advanced Research in Computer and Communication Engineering, 2023.
[10] Raunak Kandoi, Deepali Dixit, Mihul Tyagi, Raghuraj Singh Yadav, Conversational AI, Journal of Emerging Technologies and Innovative Research (JETIR), 2024.