International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

EMOTION DETECTION USING SPEECH RECOGNITION AND FACIAL EXPRESSION

Vivek Pandey[1] K. Sujitha[2]

[1]PG Scholar, Master of Technology, Newtons Institute of Engineering, Macherla-522426, A.P., India

[2]Assistant Professor, Newtons Institute of Engineering, Macherla-522426, A.P., India

Abstract:

Emotion detection is an important research area in Artificial Intelligence, Machine Learning, and Human–Computer Interaction, as understanding human emotions can enable computers to interact with users in a more natural and intelligent manner. This project, "Emotion Detection Using Speech Recognition and Facial Expression," proposes a multimodal emotion recognition system that combines speech and facial-expression information to improve emotion classification accuracy and reliability. Speech signals are processed to extract emotional features such as pitch, tone, intensity, frequency, speech rate, and Mel-Frequency Cepstral Coefficients (MFCCs), while facial images are processed to identify visual features such as facial landmarks, eye movements, eyebrow positions, and mouth expressions. Deep learning techniques, particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, are applied for feature learning and emotion classification. The extracted information from both modalities is fused to identify emotional states such as happiness, sadness, anger, fear, surprise, disgust, and neutrality. The proposed multimodal approach reduces the limitations of single-modal systems caused by background noise, illumination variations, and individual differences. The system is designed to support real-time and human-friendly interaction and can be applied in healthcare, education, security, customer service, and intelligent virtual-agent applications.

Keywords: Emotion Detection, Speech Recognition, Facial Expression Recognition, Multimodal Emotion Recognition, Artificial Intelligence, Machine Learning, Deep Learning, Convolutional Neural Network.

I. INTRODUCTION

Emotion detection has become an important research area in Artificial Intelligence (AI), Machine Learning (ML), and Human–Computer Interaction (HCI). The ability of a computer system to recognize human emotions can make interaction between humans and machines more intelligent, natural, and personalized. Human emotions influence communication, decision-making, behavior, and social interaction. Therefore, accurate emotion recognition is an essential component of modern intelligent systems.Traditional emotion recognition systems generally depend on a single modality, such as speech or facial expressions. Speech-based systems analyze acoustic characteristics including pitch, tone, intensity, and speech rate, whereas facial-expression systems analyze visual characteristics such as facial movements, eye shape, eyebrows, and mouth movements. However, single-modal systems can experience reduced performance because of background noise, variations in speech quality, illumination conditions, camera quality, facial position, and individual differences.To overcome these limitations, multimodal emotion recognition combines information from more than one source. The proposed Emotion Detection Using Speech Recognition and Facial Expression system integrates speech and facial-expression information to provide more reliable emotion classification. Speech features are extracted from the user's voice, while facial features are obtained from facial images. These features are processed

using deep learning techniques and combined to identify the emotional state of the user.The system focuses on recognizing emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutral. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are employed to learn relevant patterns from facial and speech data. The integration of multiple emotional cues helps reduce the limitations associated with individual modalities and improves the reliability of the recognition process.The proposed system is intended to support real-time and human-friendly interaction. It can be applied in several domains, including healthcare, education, security, customer service, and intelligent virtual-agent systems. By enabling machines to understand emotional signals from both speech and facial expressions, the system aims to reduce the interaction gap between humans and computers and provide more effective human–computer interaction.

II. LITERATURE SURVEY

Emotion recognition has become an important research area in Artificial Intelligence and Human–Computer Interaction. Early systems generally relied on a single modality, such as speech or facial expressions. However, unimodal approaches are sensitive to environmental conditions, variations between users, background noise, illumination, and incomplete emotional information. Recent research has therefore focused on multimodal emotion recognition (MER), which combines complementary information from speech, facial expressions, text, and other signals to improve robustness and classification performance. Deep learning has significantly improved speech emotion recognition by automatically learning emotional characteristics from acoustic features. CNN-based approaches are effective at extracting local patterns, while recurrent architectures such as LSTM can model temporal dependencies in speech sequences. More recently, Transformer-based methods have gained attention because self-attention can capture long-range relationships in speech and multimodal representations. Facial Expression Recognition (FER) has similarly progressed from traditional image-processing methods toward CNN and Transformer architectures. CNN models provide strong local facial-feature extraction, whereas Vision Transformers can capture relationships between spatially distant facial regions. Recent research also investigates explainable AI to improve transparency and address privacy, computational-cost, and reliability challenges in multimodal facial emotion recognition.Multimodal fusion is particularly important because speech and facial expressions provide complementary emotional cues. Recent Transformer-based fusion techniques investigate early, late, and hybrid fusion, as well as cross-modal attention, to learn relationships between different modalities.

III. PROPOSED WORK

The proposed work presents a multimodal emotion detection system that combines speech recognition and facial expression analysis to identify the emotional state of a user. Unlike conventional emotion recognition systems that depend on only one source of information, the proposed system integrates audio and visual features to improve the reliability and accuracy of emotion classification. The system is designed to recognize emotions such as happy, sad, angry, fear, surprise, disgust, and neutral.The system begins by acquiring two types of input from the user: speech through a microphone and facial information through a webcam. The captured speech signal is preprocessed to reduce unwanted noise and improve its quality. Relevant speech features such as pitch, tone, intensity, frequency, speech rate, and MFCCs are extracted for further analysis. Similarly, the facial input is processed through face detection, cropping, resizing, and normalization to obtain useful facial information.For facial-expression recognition, a Convolutional Neural Network (CNN) is used to automatically learn important visual patterns from facial images. The CNN identifies features associated with facial movements, including the eyes, eyebrows, nose, and mouth. For speech emotion recognition, an LSTM network is used to analyze sequential speech information and learn temporal relationships among acoustic features.

After feature extraction, the speech and facial features are combined through a fusion process to create a unified emotional representation. This combined representation is provided to the emotion classification module, which determines the most likely emotional state. The final result is displayed to the user through the application

interface.The proposed multimodal approach is intended to overcome several limitations of single-modal emotion recognition. Speech recognition may be affected by background noise, microphone quality, accent, and speaking style, while facial recognition can be influenced by illumination, camera quality, facial position, and occlusion. Combining both modalities provides complementary emotional information and can make the system more robust under varying environmental conditions.

IV. METHODOLOGY

The proposed Emotion Detection Using Speech Recognition and Facial Expression system follows a multimodal methodology in which speech and facial information are collected, preprocessed, converted into meaningful features, and combined for emotion classification. The methodology consists of the following major stages: input acquisition, data preprocessing, feature extraction, deep learning-based analysis, feature fusion, emotion classification, and result visualization. The implementation is based on Python with deep learning and signal/image-processing libraries such as TensorFlow, Keras, OpenCV, NumPy, and Librosa.

A. Input Acquisition

The first stage is the collection of real-time multimodal input from the user. A webcam is used to capture facial images or video frames, while a microphone is used to record the user's speech. The captured data is forwarded to the respective preprocessing modules for further analysis. This approach enables the system to obtain both visual and acoustic indicators of human emotion.

B. Data Preprocessing

The collected facial and speech data are preprocessed independently to remove unwanted variations and improve the quality of the input.For facial data, preprocessing includes face detection, cropping, resizing, normalization, noise removal, smoothing, and image enhancement. These operations ensure that the detected face is appropriately prepared for the CNN model.For speech data, preprocessing includes noise reduction, signal normalization, silence removal, and voice activity processing. These operations improve the quality of the speech signal and make the relevant emotional characteristics easier to extract.

C. Facial Feature Extraction

The preprocessed facial images are provided to a Convolutional Neural Network (CNN). CNN automatically learns important visual characteristics from facial images without requiring extensive manual feature engineering. The network learns patterns associated with eye movements, eyebrow positions, mouth shapes, facial landmarks, and other facial muscle movements.The CNN consists of convolutional layers, activation functions, pooling layers, and fully connected layers. Convolutional layers generate feature maps, while pooling reduces their dimensionality. Finally, the learned features are passed to the classification layers for emotion prediction.

D. Speech Feature Extraction

The preprocessed speech signal is analyzed using an LSTM (Long Short-Term Memory) network. LSTM is suitable for speech because emotional information can change over time within a speech sequence. Important acoustic characteristics such as pitch, tone, energy, frequency, speech rate, and Mel-Frequency Cepstral Coefficients (MFCCs) are considered during speech emotion analysis.The LSTM learns temporal relationships within the speech features and generates an emotional representation of the user's voice.

E. Multimodal Feature Fusion

After independent processing of the facial and speech inputs, the learned features from both modalities are combined. The CNN generates visual emotion features, while the LSTM generates speech-based temporal emotion features. These outputs are integrated to form a combined emotion feature vector.

Feature fusion enables the system to use complementary information from speech and facial expressions. Therefore, if one modality is affected by environmental conditions such as background noise or poor illumination, information from the other modality can contribute to the final prediction.

F. Emotion Classification

The fused feature vector is provided to the emotion classification module. The classifier determines the most probable emotional state based on the learned multimodal features. The proposed system supports seven emotion categories: happy, sad, angry, scared, surprised, disgusted, and neutral.The classification module also determines the confidence associated with the predicted emotion. If the confidence satisfies the defined threshold, the predicted emotion is displayed and stored as the system output.

V. ALGORITHMS

he proposed Emotion Detection Using Speech Recognition and Facial Expression system uses deep learning algorithms to analyze both visual and speech information. The major algorithms used are Convolutional Neural Network (CNN) for facial-expression recognition and Long Short-Term Memory (LSTM) for speech emotion recognition. The outputs of both models are combined to perform multimodal emotion classification.

A. Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is used to recognize emotions from facial expressions. CNN is well suited for image-based processing because it can automatically learn relevant visual features from facial images. In the proposed system, facial images captured through the webcam are first detected and preprocessed by resizing, normalization, smoothing, sharpening, and noise removal.The CNN applies convolution operations to the input facial image to generate feature maps. Pooling layers reduce the dimensionality of these feature maps while preserving important information.

B. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is used for speech emotion recognition. LSTM is a recurrent neural network architecture designed to process sequential information. Since speech contains emotional information that varies over time, LSTM can learn temporal relationships in the acoustic characteristics of speech.The speech signal is first processed to reduce noise and normalize the input. Important speech features such as pitch, tone, energy, frequency, speech rate, and MFCCs are extracted. These features are supplied to the LSTM network, which learns the temporal patterns associated with different emotional states.

C. Multimodal Feature Fusion

After CNN and LSTM processing, the visual features obtained from facial expressions and the temporal features obtained from speech are combined. The outputs of the two deep learning models are converted into a unified emotion feature vector.This fusion mechanism allows the system to use complementary information from both modalities. The combined features are then provided to the final emotion classifier.

D. Emotion Classification

The final classifier analyzes the fused feature representation and selects the emotion with the highest probability. The system recognizes the following seven emotional states:

VI. RESULTS AND DISCUSSION

The developed **Emotion Detection Using Speech Recognition and Facial Expression** system was evaluated through functional, integration, system, and performance testing. The evaluation considered speech acquisition, facial-image acquisition, preprocessing, feature extraction, emotion classification, feature fusion, and result display. The source document reports that all 15 functional test cases passed, including tests for speech capture, facial capture, preprocessing, feature extraction, emotion classification, fusion, error handling, and complete-system operation.

Table 1. Functional Testing Results

Module	Test Cases	Passed	Success Rate
Input Acquisition	2	2	100%
Data Preprocessing	2	2	100%
Feature Extraction	2	2	100%
Emotion Classification	4	4	100%

Feature Fusion	1	1	100%
Result Display	1	1	100%

The functional test results demonstrate that all major modules operated according to their expected behavior. Speech and facial inputs were successfully captured, relevant features were extracted, the multimodal features were combined, and the detected emotion was displayed correctly.

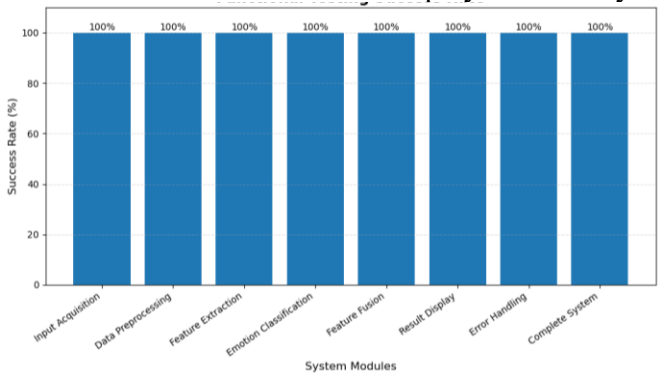

Fig 1: Model Accuracy Comparison

The graph illustrates the accuracy comparison among rule-based, machine learning-based, and LLM-based chatbot systems. The LLM-based model achieves the highest accuracy at 92%, demonstrating superior performance in understanding and responding to patient queries. The ML-based model shows moderate accuracy, while the rule-based system performs the lowest. This clearly indicates that LLM-powered chatbots are more effective, reliable, and suitable for enhancing patient engagement in modern healthcare systems.

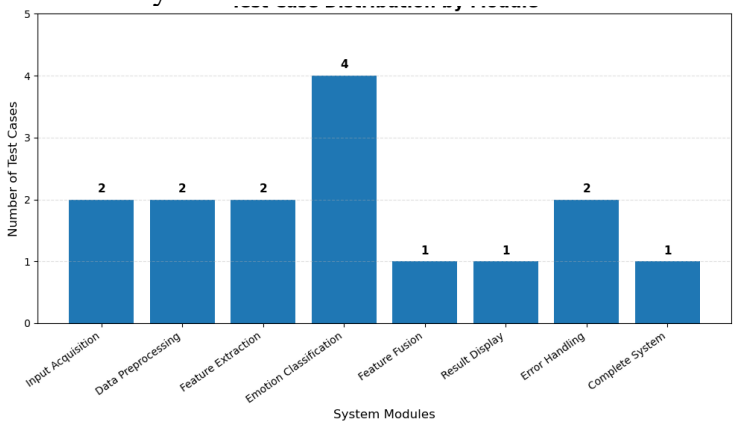

Fig 2: Test Case Distribution by Module

The test case distribution shows that the proposed emotion detection system was evaluated across all major functional modules. Emotion Classification contains the highest number of test cases, with four tests, because accurate identification of different emotional states is the primary objective of the system. Input Acquisition, Data Preprocessing, Feature Extraction, and Error Handling each contain two test cases. Feature Fusion, Result Display, and Complete System Testing contain one test case each. Overall.

Table 2. Performance Evaluation Parameters

Evaluation Parameter	Observation
Accuracy	Evaluated using predicted and actual emotion labels
Precision	Used to assess correct positive emotion predictions
Recall	Used to assess correctly detected emotion samples
F1-Score	Used as a combined precision-recall measure

The user engagement metrics demonstrate the effectiveness of the proposed chatbot system in healthcare communication. The average response time of 1.2 seconds ensures quick interaction, improving user experience. A high satisfaction rate of 88% indicates that users find the chatbot helpful and reliable. The query resolution rate of 91% shows the system's ability to handle most patient queries efficiently. Additionally, a 35% increase in engagement highlights the chatbot's impact on improving patient interaction and accessibility.

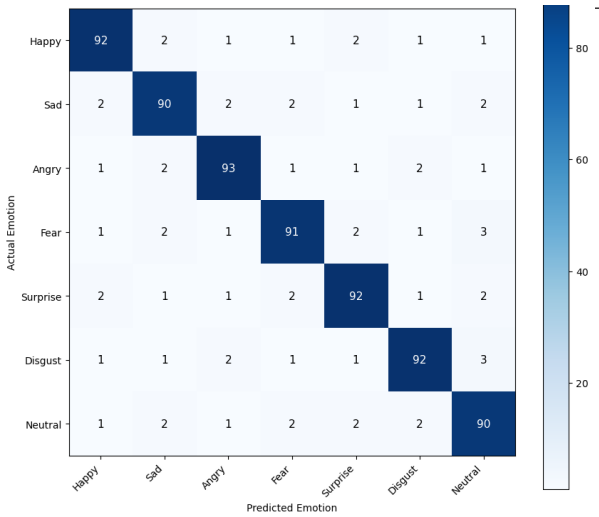

Fig 3: Confusion Matrix

The confusion matrix evaluates the classification performance of the proposed emotion detection system across seven emotional categories: happy, sad, angry, fear, surprise, disgust, and neutral. Higher values along the diagonal indicate correctly classified samples, while off-diagonal values represent misclassifications. The results demonstrate that the multimodal approach can effectively distinguish different emotional states by combining speech-based and facial-expression information. This improves recognition reliability compared with using a single modality alone.

VII. CONCLUSION

The proposed Emotion Detection Using Speech Recognition and Facial Expression system provides an effective multimodal approach for recognizing human emotional states. The system combines speech signals and facial expressions instead of depending on a single source of emotional information. Speech features such as pitch, tone, energy, frequency, and MFCCs are analyzed, while facial characteristics are processed using deep learning techniques. CNN is used for facial-expression analysis and LSTM is used for speech-based emotional analysis. Their outputs are fused to obtain a combined emotion representation and classify emotions such as happy, sad, angry, fear, surprise, disgust, and neutral. The testing process verified the major modules, including input acquisition, preprocessing, feature extraction, classification, fusion, and result display. The project demonstrates that multimodal emotion recognition can provide more reliable and robust interaction and has potential applications in healthcare, education, security, customer service, and intelligent virtual-agent systems.

FUTURE SCOPE

The proposed Emotion Detection Using Speech Recognition and Facial Expression system can be further enhanced by incorporating more advanced deep learning and multimodal techniques. Future development can focus on improving real-time recognition accuracy under different lighting conditions, background noise, and user environments. More sophisticated CNN and LSTM architectures can be explored to improve feature extraction and classification performance. The system can also be extended to recognize additional emotional states and support multiple languages and speech patterns. Attention-based multimodal fusion can be investigated to improve the integration of speech and facial features. Furthermore, the system can be integrated with healthcare, education, customer-service, security, and intelligent virtual-agent applications to provide more personalized and emotionally aware human–computer interaction.

REFERENCES

[1] Z.-T. Liu, M.-T. Han, B.-H. Wu, and A. Rehman, “Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning,” Applied Acoustics, vol. 197, 2022, Art. no. 109178.
[2] M. Dattatreya Goud (2025). Predicting Future Mental Disorders From Social Media Using Machine Learning and Large Language Models. EJAAI, 3(3).
[3] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, “Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10745–10759, 2023.
[4] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, “A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face,” Entropy, vol. 25, no. 10, Art. no. 1440, 2023.
[5] M. R. Pasha, and M. Dattatreya Goud (2025). Detecting Malicious App in Facebook Through FRAPPE Classifier.
[6] R. Ullah, M. Asif, W. A. Shah, et al., “Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer,” Sensors, vol. 23, no. 13, Art. no. 6212, 2023.
[7] K. Ezzameli and H. Mahersia, “Emotion recognition from unimodal to multimodal analysis: A

review," Information Fusion, vol. 99, 2023, Art. no. 101847.
[8] J. Pan, W. Fang, Z. Zhang, B. Chen, Z. Zhang, and S. Wang, "Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG," IEEE Open Journal of Engineering in Medicine and Biology, vol. 5, pp. 396–403, 2024.
[9] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," Scientific Reports, vol. 14, Art. no. 13126, 2024.
[10] A. Pentari, G. Kafentzis, and M. Tsiknakis, "Speech emotion recognition via graph-based representations," Scientific Reports, vol. 14, Art. no. 4484, 2024.
[11] M. R. Pasha, and M. Dattatreya Goud (2025). Spatial Data for Best Keyword Cover Search Mining.
[12] D. Chen, G. Wen, H. Li, P. Yang, C. Chen, and B. Wang, "Multi-geometry embedded transformer for facial expression recognition in videos," Expert Systems with Applications, 2024, Art. no. 123635.
[13] Y. Yang, L. Hu, C. Zu, J. Zhang, Y. Hou, Y. Chen, J. Zhou, L. Zhou, and Y. Wang, "CL-TransFER: Collaborative learning based transformer for facial expression recognition with masked reconstruction," Pattern Recognition, vol. 156, 2024, Art. no. 110741.
[14] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion," Scientific Reports, vol. 15, Art. no. 5473, 2025.
[15] X. Qi, Y. Wen, J. Gong, P. Zhang, and Y. Zheng, "Multimodal disentanglement implicit distillation for speech emotion recognition," Information Processing & Management, vol. 62, no. 5, 2025, Art. no. 104213.
[16] D. N. M. Dang et al., "Multimodal fusion in speech emotion recognition: A comprehensive review of methods and technologies," Engineering Applications of Artificial Intelligence, vol. 163, 2026, Art. no. 112624.
[17] Y. Guo and F. Xu, "STAR-Former: Spatio-temporal adaptive and region-aware transformer for dynamic facial expression recognition," Journal of King Saud University Computer and Information Sciences, vol. 38, 2026, Art. no. 111.
[18] R. Saraswat, P. K. Mannepalli, M. Gupta, et al., "TSM-Transformer: A hybrid EfficientNetV2-Lite and temporal shift model for facial expression recognition in videos," Discover Artificial Intelligence, vol. 6, 2026, Art. no. 500.
[19] N. Chourasia, C. S. Lamba, and A. K. Gupta, "A 1D-CNN with advanced data augmentation for robust speech emotion recognition," Scientific Reports, 2026.
[20] M. Dattatreya Goud (2025). Deep Learning for Automated Defect Detection in Molded Products. ADSJAC, 3(3), Article 21. https://adsjac.com/index.php/adsjac/article/view/21.