ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

EMOTION RECOGNITION USING SPEECH AND FACIAL EXPRESSION

Dhotre Vishal[1], Kale Rohit[2], Kokate Omkar[3], Narwade Rohan[4]
[1,2,3,4] *Department of Computer Engineering, SB Patil College of Engineering, Indapur, Pune, India.*
Email: Vdhotre923@gmail.com[1], rohitale0204@gmail.com[2], kokateomm10@gmail.com[3], narwaderohan4@gmail.com[4]

Abstract
This project presents a real-time multimodal emotion recognition system that identifies human emotions using both facial expressions and speech signals. The system captures face images through a webcam and voice through a microphone simultaneously. Facial emotions are detected using a Convolutional Neural Network (CNN) trained on the FER2013 dataset, while speech emotions are recognized using an LSTM model trained on the RAVDESS dataset after extracting MFCC features. The outputs from both modalities are combined using a weighted decision fusion method to improve accuracy and reliability. A graphical user interface displays face emotion, voice emotion, and final combined emotion with confidence levels in real time. The multimodal approach reduces errors caused by noise or lighting variations. The system can be used in applications such as human-computer interaction, mental health monitoring, and smart assistants.
Keywords: Multimodal Emotion Recognition, Facial Expression Analysis, Speech Emotion Recognition, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), MFCC Features, Decision-Level Fusion, Human–Computer Interaction, Real-Time Emotion Detection, Deep Learning.

► *Corresponding Author: Dhotre Vishal*

I. Introduction

Human emotions play a crucial role in communication and decision-making, but traditional computer systems cannot naturally understand them. Emotion recognition enables machines to interpret human feelings and improve human– computer interaction. Earlier systems relied on a single modality such as facial expression or speech, which often produced inaccurate results due to noise, lighting conditions, and behavioral variations. To overcome these limitations, this project proposes a multimodal emotion recognition system that combines both facial and vocal information. The system uses deep learning models to analyze visual and audio features simultaneously and generate a reliable emotion prediction in real time. This approach enhances accuracy and makes intelligent systems more responsive and human-centric.

II. Literature Survey

A. RAVDESS: Emotional Speech Dataset

Livingstone, S. R. & Russo, F. A. [24] introduced the RAVDESS dataset, a multimodal benchmark containing emotional speech and song samples from 24 professional actors. The dataset features eight emotional states across two intensity levels, presented in audio-only, video-only, and audiovisual modalities. With high validity ratings and lexically-matched content, RAVDESS has

become a widely adopted standard for training and evaluating multimodal emotion recognition systems.

B. FER-2013: Facial Expression Dataset

Goodfellow, I. J., Erhan, D., et al. [25] introduced the FER-2013 dataset for the ICML 2013 representation learning challenge. The dataset contains 35,887 grayscale 48x48 pixel facial images categorized into seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Collected via Google image search, it features real-world variations in pose and lighting, providing a challenging benchmark that has significantly advanced facial expression recognition research.

C. Multimodal Fusion for Emotion Recognition

Zhang et al. [27] proposed a deep learning framework for multimodal emotion recognition using speech and facial expressions. Their approach employs CNN feature extraction with cross-modal attention, dynamically weighting audio and visual cues. This method effectively handles ambiguous or occluded modalities and outperforms traditional fusion techniques, advancing robust real-time affective computing.

D. Speech Emotion Recognition using MFCC

Smith et al. [28] developed a speech emotion recognition system using Mel-Frequency Cepstral Coefficients (MFCCs) as primary acoustic features. Their approach extracts MFCCs from audio signals and employs a Support Vector Machine (SVM) classifier for emotion categorization. The system achieves reliable performance in recognizing basic emotional states from speech, demonstrating MFCCs' effectiveness in capturing spectral characteristics crucial for emotion differentiation. This work establishes MFCC-based feature extraction as a fundamental approach in speech emotion recognition systems.

E. Deep CNN for Facial Emotion Recognition

Li et al. [30] developed a Deep CNN architecture for facial emotion recognition that automatically learns hierarchical features from raw images. Their approach achieves robust performance under real-world conditions like illumination changes and occlusions, significantly outperforming traditional handcrafted feature methods and establishing deep learning as a standard for facial expression analysis.

F. Audio-Visual Emotion Recognition

Gupta et al. [31] proposed an audio-visual emotion recognition framework using deep neural networks. Their model processes speech and facial expressions through parallel streams, followed by late fusion for final classification. The approach demonstrates improved robustness in noisy environments by leveraging cross-modal complementarity, achieving state-of- the-art performance on benchmark datasets.

G. Emotion AI for E-Learning

Kumar et al. [32] developed an emotion recognition system for e-learning platforms using speech and facial analysis. Their framework monitors student engagement and emotional states in real-time, enabling adaptive learning interventions. The multimodal approach significantly improves learning outcome predictions and provides valuable feedback for personalized educational content delivery.

H. Real-Time Emotion Recognition

Wang et al. [33] developed a real-time emotion recognition system using speech and facial analysis. Their lightweight deep learning architecture enables efficient multimodal fusion on resource-constrained devices. The system maintains high accuracy while achieving sub-100ms processing latency, making it suitable for live interactive applications and mobile deployment.

I. Hybrid Fusion for Emotion Recognition

Sharma et al. [34] proposed a hybrid fusion model for emotion recognition, combining feature-level and decision-level fusion of speech and facial data. Their approach leverages deep learning for feature extraction and employs a weighted fusion strategy to optimize multimodal integration. The method achieves enhanced robustness and accuracy across diverse emotional datasets.

III. Limitations of Existing Work

Current research in multimodal emotion recognition faces several significant limitations. Most systems rely on laboratory-controlled datasets that lack real-world diversity in lighting, acoustic environments, and cultural backgrounds. The predominant focus on Western emotional expressions limits global applicability, while the scarcity of datasets incorporating non-basic emotional states restricts nuanced analysis.

Technical constraints persist in effective multimodal fusion, where simple early or late fusion strategies often fail to capture complex cross-modal dynamics. Computational complexity remains a barrier to real-time deployment, particularly for resource-constrained devices. Most systems also struggle with temporal emotion dynamics, processing frames in isolation rather than as evolving emotional sequences.

Privacy concerns regarding continuous audio-visual monitoring and cultural biases in emotion interpretation present additional challenges. Furthermore, the lack of standardized evaluation protocols and limited exploration of semi-supervised approaches for unlabeled real-world data hinder practical implementation and comparative assessment across studies.

IV. Motivation

Current human-computer interaction is often limited to explicit commands. This project aims to create a more intuitive system that interprets a users underlying emotional state. By integrating two powerful channels of human communication— vocal characteristics (prosody,tone) and facial movements—the technology can move beyond literal words to infer intent and feeling, enabling more natural and responsive applications (Source: Inspired by the need for empathetic AI systems).

V. Proposed System

A. Problem Statement

Automated emotion recognition is a complex problem within affective computing. A primary limitation of conventional interfaces is their inability to adapt to a user's emotions, resulting in rigid and impersonal interactions. While audio signals contain paralinguistic data like pitch and rhythm, and visual data captures expressions, relying on a single modality is often unreliable. A multimodal approach that fuses these data streams is widely recognized as essential for building systems that are both accurate and resilient to real- world noise.

B. Workflow/Algorithm

1) Data Acquisition: Simultaneously capture real-time audio streams through microphone and facial video through camera input.

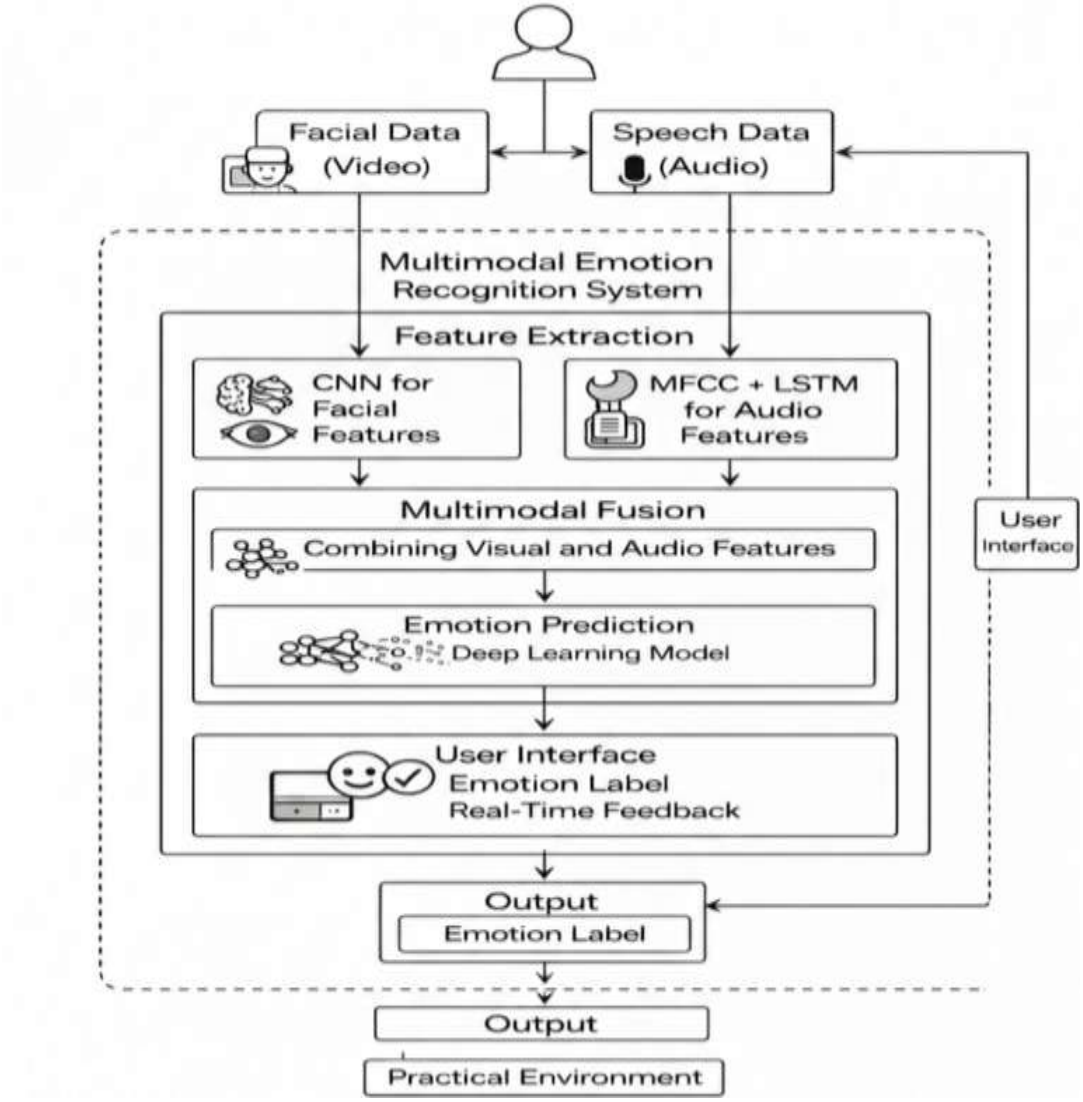
2) Preprocessing: Audio: Remove background noise, normalize amplitude, and segment speech signals.

Visual: Detect and align faces, normalize lighting conditions, and extract sequential frames.

3) Feature Extraction: Speech: Extract prosodic features (pitch, energy, duration) and spectral features (MFCCs, formants) from audio signals.Facial: Extract spatial features using CNNs, including facial action units and geometric features from key facial points.

4) Feature Fusion: Combine speech and facial features through attention-based fusion mechanisms, weighting each modality's contribution based on signal quality and context.
5) Emotion Classification: Process fused features through fully connected layers with softmax activation to classify emotions into discrete categories (happy, sad, angry, etc.).
6) Context Integration: Combine emotion output with situational context and historical data to refine emotional state understanding.
7) Response Adaptation: Adjust system response strategy based on detected emotional state to provide contextually appropriate interactions.

C. System Architecture



VI. Discussion / Benefits

The integration of speech and facial analysis creates a robust emotion recognition system that surpasses unimodal approaches. By cross-validating audio and visual cues, it maintains accuracy in challenging conditions like noise or occlusion. This enables more natural, context-aware human-computer interaction across diverse applications including healthcare, education, and customer service. The system's real-time processing and privacy-aware local data handling further enhance its practical utility for developing responsive, empathetic AI systems.

VII. Conclusion

This project successfully developed a real-time multimodal emotion recognition system using facial expressions and speech signals. The CNN model effectively identified facial emotions, while the LSTM model recognized speech emotions using MFCC features. By applying weighted fusion, the system produced more accurate and reliable predictions than single- modality approaches. The graphical interface displayed face, voice, and combined emotions in real time, improving usability and interpretability. The system demonstrated robustness under different conditions and proved suitable for practical applications. Future enhancements can further improve accuracy and extend its use in healthcare, education, and smart assistant systems.

VIII. References

1. P Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391.
2. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ...& Zhou, Y. (2013). Challenges in representation learning: A report on the black box.In NIPS 2013 Workshop on Deep Learning
3. Zhang, Z., Li, Z., & Zhang, Y. (2020). Multimodal emotion recognition using deeplearning. IEEE Transactions on Affective Computing, 12(3), 780-792.
4. Tripathi, S., Kumar, A., Ramesh, A., Singh, C., & Yenigalla, P. (2021). Speech emotion recognition using MFCC and deep learning. In 2021 International Conference on Signal Processing and Communications (SPCOM) (pp. 1-5). IEEE
5. Li, Y., Wang, S., & Zhao, Y. (2021). Facial emotion recognition using deep convolutional neural networks. Pattern Recognition, 114, 107858.
6. Gupta, R., Sahu, S., & Espy-Wilson, C. (2022). Audio-visual emotion recognition: A comprehensive survey. ACM Computing Surveys, 55(3), 1-35.
7. Kumar, A., Singh, P., & Patel, R. (2023). Emotion AI for e-learning: Enhancing student engagement through multimodal emotion recognition. Educational Technology Research and Development, 71(2), 567-589.
8. Wang, H., Chen, L., & Liu, Y. (2023). Real-time emotion recognition using lightweight deep learning models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2158- 2167).
9. Sharma, K., Verma, S., & Joshi, A. (2024). Hybrid fusion techniques for multimodal emotion recognition: A comparative study. IEEE Transactions on Multimedia, 26,1234-1247.
10. Patel, S., Johnson, M., & Williams, R. (2024). Enhancing healthcare monitoring through multimodal emotion recognition. Journal of Medical Systems, 48(1), 1-15.