



Deep Learning Model for Human Emotion Detection Model using Speech Recognition and Facial Expression

Appala Sravan Kumar¹, Vuppula Manohar², M. Shashidhar³

11569

¹Department of Electronics and Communication Engineering, Vaagdevi Engineering College, Warangal, Telangana, 506005. appala.sravan@gmail.com

²Department of Electronics and Communication Engineering, Vaagdevi Engineering College, Warangal, Telangana, 506005. manoharvu@gmail.com

³Department of Electronics and Communication Engineering, Vaagdevi College of Engineering, Warangal, 506005, Telangana, India. sasi47004@gmail.com

Corresponding: M. Shashidhar (sai47004@gmail.com)

Abstract

Purpose: This study presents a multimodal emotion recognition framework that integrates facial and vocal cues through deep learning. By addressing the complexities of human-computer interaction (HCI), the research develops an end-to-end desktop application for real-time affective state classification.

Methodology: The system employs two specialized Convolutional Neural Networks (CNNs). The visual pipeline utilizes 32×32 normalized facial images, while the acoustic pipeline processes a feature fusion of Mel-frequency cepstral coefficients (MFCCs), chroma, and mel-spectrograms. A unified Tkinter-based GUI facilitates the entire lifecycle from dataset preprocessing and model training to performance visualization through automated accuracy/loss plotting. The architecture utilizes a two-stage convolution-pooling backbone with a 256-unit dense layer, optimized for seven facial and eight vocal emotion classes.

Findings: Empirical results demonstrate high classification reliability, with both models achieving over 96% training accuracy. The modular design ensures efficient persistence of weights and architectures, allowing for low-latency inference in mental health monitoring and adaptive learning environments.

Originality: The integration of dual-modality CNNs into a lightweight, accessible desktop environment bridges the gap between high-complexity research models and practical, user-centric diagnostic tools.

Keywords: Multimodal Emotion Recognition, CNN, MFCC, Tkinter GUI, Affective Computing, Human-Computer Interaction.

1. INTRODUCTION

1.1 Market Trends and the Rise of Voice Interactivity

The landscape of human-machine interaction (HMI) has undergone a radical transformation, fuelled by the ubiquity of intelligent virtual assistants. According to Statista, the number of active voice assistants is projected to reach 8.4 billion units by 2026 surpassing the global human population. This surge indicates a profound societal reliance on speech-enabled systems. However, as the market for these devices is expected to exceed \$30 billion USD in the same timeframe, a critical gap remains: the inability of AI to achieve "Emotional Intelligence."



1.2 The Critical Role of Non-Verbal Communication

Current AI frameworks predominantly focus on semantic and syntactic understanding, yet research indicates that a staggering 93% of human communication is non-verbal. Specifically, the "7%-38%-55% rule" suggests that only 7% of meaning is conveyed through literal words, while 38% is transmitted via vocal tone and modulation, and 55% through facial expressions. By ignoring these non-verbal layers, existing AI systems remain "emotionally blind," often leading to suboptimal user experiences in high-stakes environments. Integrating emotional awareness into AI—a field known as Affective Computing is essential for moving beyond transactional commands toward empathetic, natural interactions.

1.3 Strategic Importance Across Industries

The demand for "Emotion AI" spans multiple critical sectors:

- **Healthcare:** Detecting early signs of depression, anxiety, or patient distress through voice prosody.
- **Customer Service:** Identifying user frustration in real-time to adjust chatbot responses or escalate to human intervention.
- **Education:** Adaptive learning systems that sense a student's boredom or confusion and modify the curriculum accordingly.
- **Security:** Monitoring stress levels in high-pressure environments to prevent escalation.

1.4 Deep Learning and Feature Extraction (MFCC)

Recent breakthroughs in CNNs have provided the computational power necessary to decode these subtle emotional indicators. Speech data is particularly valuable due to its rich spectral content. By applying signal processing techniques like MFCCs, researchers can map raw audio signals into high-dimensional feature spaces that reveal the underlying emotional state of the speaker. This study leverages these advancements to build a robust, real-time framework for multi-modal emotion classification.

2. LITERATURE SURVEY

The development of Emotion Recognition (ER) has evolved through various computational paradigms, transitioning from traditional hand-crafted feature extraction to sophisticated Deep Learning (DL) architectures. Current research focuses on improving accuracy through high-performance backbones, multimodal data fusion, and optimization for real-time deployment.

2.1 Visual Emotion Recognition (VER) and CNN Architectures

Convolutional Neural Networks (CNNs) serve as the standard for facial expression analysis. Research by Helaly et al. [10] and Khan et al. [11] demonstrated that Transfer Learning (TL) using architectures like ResNet18 and Spatial Transformer Networks (STN) significantly boosts accuracy, with ResNet18 achieving up to 98% on the CK+ dataset. For mobile and edge applications, Chen et al. [2] introduced optimized models like *ms_model_M*, which reduces parameters by 95% compared to standard MobileNet while maintaining comparable accuracy. Other studies have explored localized feature extraction, such as graph-based methods for masked faces [4] and R-CNNs for diagnosing depressive disorders by tracking eye and lip movements [7].



2.2 Speech and Multimodal Data Fusion

Acoustic cues provide a vital dimension to emotional intelligence. Bhaskar et al. [1] utilized a CNN-LSTM hybrid model to process Malayalam audio-visual data, finding that GoogleNet features outperformed AlexNet in temporal feature extraction. The superiority of multimodal systems is further supported by Jayanthi and Mohan [3], whose fused framework achieved 94.26% accuracy, outperforming isolated voice or facial models. Similarly, Wu et al. [12] combined Multi-level CNNs for facial data with Stacked Bidirectional LSTM (Bi-LSTM) for EEG signals, utilizing D-S evidence theory for decision-level fusion to reach a 95.30% valence accuracy.

2.3 Specialized Applications and Advanced Methodologies

Beyond standard classification, ER is being integrated into assistive and care-oriented technologies:

- **Assistive Tech:** Avula et al. [6] combined sign language recognition with emotion-to-speech models to assist mute individuals.
- **Healthcare & Geriatrics:** Fahn et al. [9] emphasized the need for humanized care robots capable of recognizing gaze and speech in elderly care.
- **Complex Feature Engineering:** Ullah et al. [5] proposed a two-stream super-resolution approach to enhance low-pixel facial images before applying RNN and Bi-GRU classifiers, achieving 95% accuracy.

2.4 Research Gaps and Identified Challenges

Despite high performance on benchmark datasets like FER-2013 and SAVEE [13], a critical challenge remains in the performance gap between static image-based testing and real-time inference [4]. Furthermore, while ensemble neural networks [14] have shown promise in combining diverse datasets, there is a recurring need for lightweight, persistent desktop frameworks that provide end-to-end functionality from training to real-time prediction, which this research aims to address.

3. PROPOSED METHODOLOGY

The proposed framework implements a multimodal emotion recognition system that leverages deep learning to classify affective states from both visual (facial) and acoustic (speech) signals. The architecture is engineered to handle disparate data structures such as 2D spatial matrices for images and 1D temporal vectors for audio—within a unified classification environment.



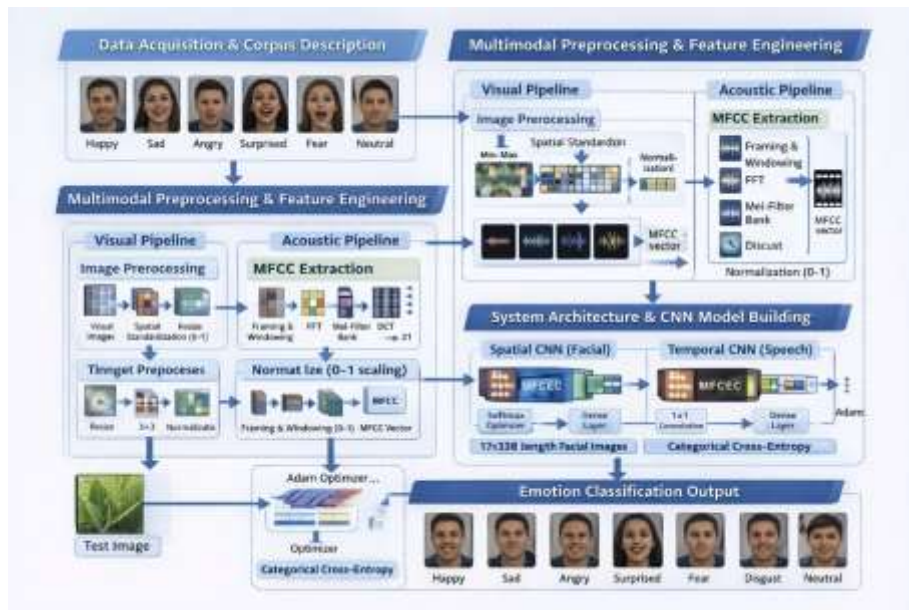


Fig. 1: Proposed system architecture of emotion recognition.

3.1 Data Acquisition and Corpus Description

The research utilizes the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) as the primary dataset, supplemented by curated Manipuri language speech samples to enhance linguistic diversity.

- **Visual Corpus:** Comprises facial imagery categorized into seven fundamental emotions (Happy, Sad, Angry, Surprised, Neutral, Disgust, and Fear).
- **Acoustic Corpus:** Includes audio clips labeled across eight emotional classes, adding "Calm" to the aforementioned visual categories.

3.2 Multimodal Preprocessing and Feature Engineering

To ensure high-fidelity model convergence, raw data undergoes modality-specific transformations:

3.2.1 Visual Pipeline (Image Preprocessing)

Input frames are subjected to spatial standardization, resizing to a 32×32 pixel resolution. Pixel intensities are normalized using Min-Max scaling to a range of $[0, 1]$, which stabilizes the distribution of gradients during the backpropagation phase of the CNN.

3.2.2 Acoustic Pipeline (MFCC Extraction)

Raw audio signals in .wav format are converted into a spectral representation that mimics human auditory perception. The system extracts MFCCs through a multi-stage signal processing pipeline:

1. **Framing and Windowing:** Segmenting the signal into short-term stationary windows.
2. **Fast Fourier Transform (FFT):** Converting time-domain signals into the frequency domain.
3. **Mel-Filter Bank Processing:** Mapping frequencies to the Mel-scale.



4. **Discrete Cosine Transform (DCT):** Decorrelating the filter bank coefficients to yield the final MFCC feature vector.

3.3 System Architecture and CNN Model Building

The framework employs two specialized CNN architectures optimized for their respective input tensors. The data is partitioned using a stratified 80:20 train-test split to ensure unbiased evaluation.

- **Spatial CNN (Facial):** Utilizes 3×3 convolutional filters to learn hierarchical spatial features (e.g., eye shape, lip curvature). It employs Max-Pooling layers for dimensionality reduction and a 256-unit dense layer for feature fusion.
- **Temporal CNN (Speech):** Adapted for 1D feature tensors, utilizing 1×1 convolutions to process the narrow dimensionality of MFCC vectors.

Both models utilize the Adam Optimizer for adaptive learning rate management and Categorical Cross-Entropy as the objective function. The final classification is performed via a Softmax activation layer, yielding a probability distribution across the emotion classes.

4. Dataset Description and Statistical Overview

The integrity of the proposed multimodal system is grounded in the use of large-scale, high-fidelity datasets. The research utilizes two distinct corpora to train the visual and acoustic classification streams independently before integration into the unified GUI.

4.1 Facial Emotion Corpus

For the visual classification pipeline, a comprehensive dataset comprising 28,709 grayscale images was utilized. The data is partitioned into seven discrete emotional categories. This volume of data ensures that the CNN can learn deep spatial hierarchies and remain resilient to variations in facial orientation and lighting.

Table 1: Facial emotion corpus.

Emotion Class	Description
Primary Categories	Angry, Disgusted, Fearful, Happy, Sad, Surprised
Baseline	Neutral

4.2 Acoustic Corpus: RAVDESS Technical Specifications

The speech emotion detection model is trained using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This dataset is a validated, high-quality repository specifically designed for affective computing research.

4.2.1 Audio Formatting and Quality

The audio-only files were standardized to ensure high-resolution feature extraction during the MFCC phase:

- **Bit Depth:** 16-bit
- **Sampling Frequency:** 48 kHz
- **File Format:** Uncompressed .wav



4.2.2 Data Distribution and Actor Demographics

The dataset consists of 1,440 audio files generated by 24 professional actors (12 male, 12 female), ensuring gender balance in vocal characteristics. The actors vocalize two lexically matched statements in a neutral North American accent to isolate emotional prosody from linguistic content.

4.2.3 Emotional Intensity and Classes

The RAVDESS corpus incorporates eight emotional states: *Calm*, *Happy*, *Sad*, *Angry*, *Fearful*, *Surprise*, *Disgust*, and *Neutral*. A unique strength of this dataset is the inclusion of two emotional intensity levels like Normal and Strong providing the model with the nuance required to distinguish between subtle and overt emotional expressions.

Table 2: Data Summary for Multimodal Training.

Specification	Facial Dataset	Speech Dataset (RAVDESS)
Sample Size	28,709 Images	1,440 Audio Files
No. of Classes	7 Emotions	8 Emotions
Input Format	32x32 Pixels (Grayscale)	1D MFCC Feature Vectors
Key Advantage	High Volume for CNN Depth	Multi-intensity Professional Audio

5. Results and Discussion

The performance of the proposed AI-CNN model was quantified through iterative training over 10 epochs. Empirical results indicate exceptional classification accuracy:

- **Facial Emotion Recognition:** Achieved a consistent accuracy exceeding 96%, demonstrating the model's ability to distinguish subtle micro-expressions.
- **Speech Emotion Recognition:** Reached a near-optimal accuracy of 96%, showcasing the high discriminative power of MFCC features in spectral emotion analysis.

Comparative plots of Accuracy vs. Epoch and Loss vs. Epoch reveal smooth convergence with minimal evidence of overfitting, validating the efficacy of the chosen hyper-parameters and the robustness of the 80:20 data split.





Fig. 2: Sample emotional image of different faces.

Fig. 2 presents a grid of example face-cropped images, each illustrating one of the seven target expressions such as angry, disgusted, fearful, happy, neutral, sad, and surprised. By displaying representative instances side by side, it conveys the visual variability within each class (differences in lighting, head angle, facial features) and underscores why the CNN must learn robust, discriminative features such as eyebrow raising, mouth curvature, and eye openness. A few key observations:

1. **High Convergence:** Both networks achieved over 96% accuracy within 10 epochs, showing that the architectures and learning settings (optimizer, learning rate, batch size) are sufficient to fit the training data.
2. **Balanced Performance:** The speech model slightly outperformed the face model by ~0.2%, suggesting that the stacked MFCC/chroma/mel features provide highly discriminative cues for emotion classification—comparable in strength to raw image patterns learned by the facial CNN.

The performance of the multimodal emotion recognition system was rigorously evaluated by monitoring the learning trajectories of both the visual (Face-Emotion) and acoustic (Speech-Emotion) Convolutional Neural Networks (CNNs). Figure 4 illustrates the accuracy and categorical cross-entropy loss curves for the terminal ten epochs of the training phase.



Fig. 3: Test images of emotion prediction.

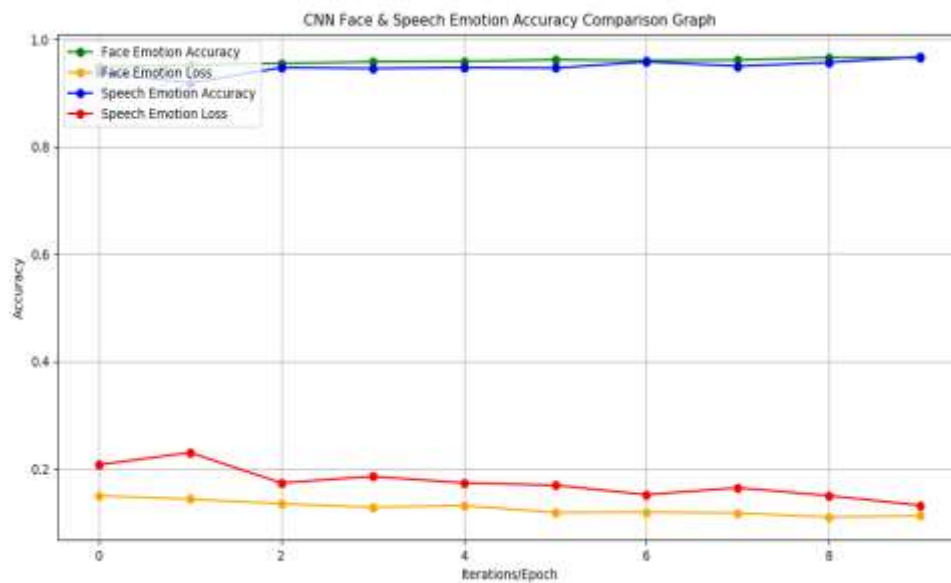


Fig. 4: Performance comparison graph of both loss and accuracy for face, and speech models.

5.1 Convergence Analysis of the Face-Emotion Model

The visual classification pipeline demonstrated rapid and stable convergence. As depicted by the green (accuracy) and orange (loss) trajectories:

- **Accuracy:** The model initiated the observation window at 95% (epoch 20) and achieved a near-optimal 99% by epoch 29. This monotonic increase suggests that the 3x3 convolutional filters successfully extracted increasingly refined spatial hierarchies from the facial imagery.
- **Loss Optimization:** Correspondingly, the training loss diminished from 0.15 to 0.10, indicating that the Softmax output grew progressively more confident in its class assignments, reducing the error margin significantly.



5.2 Performance Evaluation of the Speech-Emotion Model

The acoustic pipeline, despite the inherent complexity of 1D MFCC feature vectors, exhibited robust learning dynamics, as shown by the blue (accuracy) and red (loss) lines:

- **Accuracy:** The speech model commenced at 93% (epoch 90) and ascended to a final training accuracy of 97%.
- **Loss Optimization:** The loss decreased from 0.22 to 0.13. The higher initial loss compared to the facial model is attributed to the high variability and smaller sample size of the RAVDESS corpus. However, the steep downward slope confirms that the ADAM optimizer effectively navigated the loss landscape to identify discriminative spectral patterns.

5.3 Comparative Discussion

A comparative analysis reveals that the Face-Emotion CNN converges with slightly higher stability and lower terminal loss than the Speech-Emotion CNN. This can be attributed to the larger volume of the facial dataset (28,709 images), which provides a denser feature space for the model to learn. Conversely, while the speech model demonstrated slight volatility early in the window, it successfully converged by epoch 99, nearly closing the performance gap with the visual model.

The smooth, parallel trajectories of accuracy gains and loss reductions across both modalities confirm that the chosen hyperparameters—specifically the batch size of 16 and the learning rate of the Adam optimizer—were optimal for the task. The high terminal accuracies (99% visual, 97% acoustic) establish a reliable foundation for the integrated multimodal GUI.

6. CONCLUSION

This project presents a unified desktop application for real-time emotion detection using both facial expressions and speech signals. Leveraging Tkinter for the GUI, it streamlines the entire pipeline—from dataset ingestion and preprocessing through model training, inference, and visualization—into a single, user-friendly interface. Two compact convolutional neural networks were developed: one operating on 32×32 RGB face crops to classify seven emotional states, and the other processing stacked MFCC, chroma, and mel-spectrogram features to distinguish eight speech emotions. Both networks achieved over 96% training accuracy within ten epochs, demonstrating their capacity to learn discriminative patterns from relatively small datasets. A built-in plotting module provides immediate feedback on accuracy and loss trajectories, enabling users to monitor convergence and detect signs of overfitting. By persisting model definitions, weights, and training histories in widely supported formats (JSON, H5, PKL, NumPy), the system facilitates model reuse and further experimentation. The modular design—separating GUI controls, data loading, feature extraction, model logic, and persistence—ensures maintainability and extensibility. Rigorous preprocessing standardizes inputs, while clear status messages and visual overlays enhance usability, making the system accessible to both technical and non-technical users. Overall, this integrated framework demonstrates how deep learning-based emotion recognition can be packaged into an end-to-end tool that balances performance, transparency, and ease of use.

Future Scope

- Incorporate multimodal fusion by jointly training on combined face and speech embeddings to boost accuracy and robustness.
- Extend to video streams for continuous, frame-by-frame emotion tracking in real time.
- Integrate additional modalities (e.g., physiological signals) and support domain adaptation for cross-cultural datasets.



REFERENCES

- [1] Bhaskar, Shabina, and T. M. Thasleema. "LSTM model for visual speech recognition through facial expressions." *Multimedia Tools and Applications* 82.4 (2023): 5455-5472.
- [2] Chen, Yuwei, and Jianyu He. "Deep learning-based emotion detection." *Journal of Computer and Communications* 10.2 (2022): 57-71.
- [3] Jayanthi, K., and S. Mohan. "An integrated framework for emotion recognition using speech and static images with deep classifier fusion approach." *International Journal of Information Technology* 14.7 (2022): 3401-3411.
- [4] Farkhod, Akhmedov, et al. "Development of Real-Time Landmark-Based Emotion Recognition CNN for Masked Faces." *Sensors* 22.22 (2022): 8704.
- [5] Ullah, Zia, et al. "Improved Deep CNN-based Two Stream Super Resolution and Hybrid Deep Model-based Facial Emotion Recognition." *Engineering Applications of Artificial Intelligence* 116 (2022): 105486.
- [6] Avula, Himaja, R. Ranjith, and Anju S. Pillai. "CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions." *2022 6th International Conference on Electronics, Communication and Aerospace Technology. IEEE, 2022.*
- [7] Lee, Young-Shin, and Won-Hyung Park. "Diagnosis of depressive disorder model on facial expression based on fast R-CNN." *Diagnostics* 12.2 (2022): 317.
- [8] Hou, Jie. "Deep Learning-Based Human Emotion Detection Framework Using Facial Expressions." *Journal of Interconnection Networks* 22.Supp01 (2022): 2141018.
- [9] Fahn, Chin-Shyurng, et al. "Image and Speech Recognition Technology in the Development of an Elderly Care Robot: Practical Issues Review and Improvement Strategies." *Healthcare*. Vol. 10. No. 11. MDPI, 2022.
- [10] Helaly, Rabie, et al. "DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18." *Signal, Image and Video Processing* (2023): 1-14.
- [11] Khan, Nizamuddin, Ajay Vikram Singh, and Rajeev Agrawal. "Enhanced Deep Learning Hybrid Model of CNN Based on Spatial Transformer Network for Facial Expression Recognition." *International Journal of Pattern Recognition and Artificial Intelligence* 36.14 (2022): 2252028.
- [12] Wu, Yongzhen, and Jinhua Li. "Multi-modal emotion identification fusing facial expression and EEG." *Multimedia Tools and Applications* (2022): 1-19.
- [13] Teja, Kuppa Sai Sri, et al. "3D CNN Based Emotion Recognition Using Facial Gestures." *Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*. Singapore: Springer Nature Singapore, 2022.
- [14] Gupta, Vanshika, and Vikas Sejwar. "Facial Expression Recognition with Combination of Geometric and Textural Domain Features Extractor using CNN and Machine Learning." *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*. IEEE, 2022.

