Speech Emotion Recognition

Anurag Shekhar[1], Vikram Chaudhari[2], Deep Patil[3], Sami Shaikh[4], Shivam Padwal[5]
[1]*Lecturer, Alamuri ratnamala Institute of Engineering and Technology*
[2,3,4,5] *Student, Alamuri ratnamala Institute of Engineering and Technology*

Abstract- ***This project aims to develop a comprehensive Face and Speech Emotion Recognition system that leverages deep learning techniques to analyze and classify human emotions through facial expressions and speech signals. The facial emotion recognition module uses convolution neural networks (CNNs) to detect and interpret facial features, enabling the system to identify a range of emotions in real-time. The speech emotion recognition module applies recurrent neural networks (RNNs) or long short-term memory (LSTM) networks to analyze vocal patterns and capture emotion-related features, such as pitch, tone, and intensity. By integrating these two modalities, the system ensures more accurate emotion detection by cross-referencing visual and auditory cues. This multimodal approach enables the system to perform effectively across diverse environments and scenarios.***

The system is designed to be applicable in various domains, such as healthcare for mental health monitoring, customer support for sentiment analysis, education for personalized learning, and human-computer interaction for adaptive interfaces. Special attention is given to data preprocessing and feature extraction to optimize model performance and reduce noise from input data. The system also addresses critical concerns related to privacy and bias by ensuring secure data handling and fairness in emotion classification. Overall, this project contributes to advancing emotion-aware artificial intelligence systems, fostering more empathetic and intuitive interactions between humans and machines in real-world applications.

Keywords - Speech Emotion Recognition (SER), Machine Learning, Deep Learning, Emotional Speech Classification

I. INTRODUCTION

As technology continues to evolve, one of the most exciting advancements is the integration of emotional intelligence into artificial intelligence (AI) systems. Understanding emotions is a fundamental part of human communication, and now, AI is beginning to recognize and interpret emotions, especially through speech. Speech Emotion Recognition (SER) is a technology that enables machines to detect emotions in human speech by analyzing characteristics like pitch, tone, speed, and volume. This ability to interpret emotions opens up new possibilities for more meaningful and responsive interactions between humans and machines. Speech, being a natural and rich medium of expression, can carry a lot of emotional information. For example, a person's voice might speed up or become sharper when they are excited or angry, while it might slow down or soften when they are sad or tired. By analyzing these vocal cues, SER systems can determine a speaker's emotional state, which allows machines to react in ways that are more in tune with human feelings. This is important in areas like virtual assistants, customer support, and even healthcare, where emotional understanding can greatly enhance the user experience[1].

One exciting area in this field is Affective Computing, also known as ACE, which is all about creating systems that can recognize and respond to human emotions. ACE goes beyond just detecting emotions – it's about making machines that can respond empathetically, much like humans would in a conversation. This makes interactions with technology feel more natural, supportive, and personal. For instance, a chatbot might adjust its tone if it detects frustration in a user's voice, or a digital assistant might offer a more comforting response if it hears sadness in the user's words. This project focuses on the intersection of Speech Emotion Recognition and Affective Computing. The goal is to explore how AI can understand emotions from speech and then use that understanding to interact in a more emotionally aware way. By using machine learning to analyze speech features and emotions, the system can learn to detect various emotional states, such as happiness, anger, or sadness. More importantly, it will also learn how to adjust its responses based on these emotions, creating more empathetic and engaging interactions.

The potential uses of this technology are vast. In customer service, AI-powered assistants could provide more personalized help by recognizing

when a customer is upset or frustrated, responding with a tone of empathy and understanding. In healthcare, emotional recognition can support doctors and therapists by offering insights into a patient's emotional state, which might be difficult to assess otherwise. In education, emotional AI could help teachers understand their students' moods and needs, improving the learning experience. Similarly, in entertainment or virtual reality, machines could create more immersive experiences by responding to the emotions of the users[2].

2.OBJECTIVE OF PROJECT

The primary objective of this project is to develop an advanced Face and Speech Emotion Recognition system that can accurately analyze human emotions using deep learning techniques. By integrating facial expression analysis and speech signal processing, the system aims to create a more reliable and precise emotion detection framework. The facial recognition module utilizes Convolutional Neural Networks (CNNs) to detect facial features and interpret expressions, while the speech recognition module employs Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to analyze vocal patterns such as tone, pitch, and intensity. By combining both modalities, the system ensures a more robust and accurate emotion classification, reducing errors that may arise from using a single mode of analysis.

Another key objective is to ensure that the system is versatile and adaptable across various real-world applications. Emotion recognition has significant applications in mental health monitoring, where it can help detect emotional distress and provide early warnings for psychological conditions such as anxiety and depression. In customer service and sentiment analysis, businesses can use this technology to understand customer emotions and improve interactions. Additionally, the system can be used in education for personalized learning, allowing teachers and e-learning platforms to adapt content based on student engagement and emotions. Furthermore, in human-computer interaction, the system can enhance user experiences by enabling machines to respond more naturally to human emotions, making interactions more intuitive and engaging.

Beyond its functionality, the project also focuses on addressing critical concerns related to data privacy, fairness, and performance optimization. Special emphasis is placed on data preprocessing and feature extraction to minimize noise and enhance model efficiency. The system is designed to be fair and unbiased, ensuring that it does not misclassify emotions based on factors like ethnicity, gender, or speech variations. Additionally, strong data security measures are implemented to protect user information and ensure ethical usage of the technology. By achieving these objectives, this project contributes to the advancement of emotion-aware AI systems, promoting a future where machines can understand and respond to human emotions in a more natural and empathetic manner. Furthermore, this project aims to push the boundaries of human-machine interaction by creating an AI system that is not only accurate but also adaptive and responsive to real-world emotional variations. Traditional emotion recognition systems often struggle with environmental noise, variations in speech, and diverse facial expressions, making real-time emotion detection a challenge. This system is designed to overcome these limitations by employing advanced deep learning techniques and a multimodal approach that enhances accuracy even in complex situations. By continuously learning and improving through machine learning algorithms, the system can adapt to different users, accents, and emotional expressions, making it highly effective in diverse environments[3].

In recent years, emotion recognition has garnered significant attention due to its potential to revolutionize human-computer interactions. As technology advances, the demand for systems that can understand and respond to human emotions in real-time has increased. These systems have the potential to be used in various fields, including healthcare, customer service, education, and entertainment. The ability to accurately recognize emotions can lead to more personalized and adaptive responses, enhancing user experiences across different platforms.

Challenges in emotion recognition**:** Emotion recognition systems face several challenges, particularly when relying on a single modality. Facial expression-based systems can be affected by factors such as lighting conditions, occlusions (e.g., glasses, masks), and variations in facial features across different individuals. Similarly, speech-based emotion recognition can be hindered by background

noise, variations in accent and pronunciation, and the speaker's emotional neutrality. These challenges necessitate a multi-modal approach that combines both facial and speech cues to provide a more holistic understanding of emotional states.

Proposed System**:** The proposed system in this project aims to address these challenges by integrating facial expression analysis and speech signal processing. By employing convolutional neural networks (CNNs) for facial recognition and recurrent neural networks (RNNs) for speech analysis, the system leverages the strengths of both modalities. CNNs are well-suited for extracting spatial features from images, while RNNs excel at capturing temporal patterns in sequential data such as speech.

Data Processing and Feature Extraction**:** In the facial emotion recognition component, the system preprocesses images by normalizing and resizing them to ensure consistency. The CNN model is then used to detect and analyze facial landmarks, such as the eyes, eyebrows, mouth, and nose, which are crucial for identifying emotional expressions. In the speech emotion recognition component, audio signals are segmented and transformed into features like Mel-frequency cepstral coefficients (MFCCs), which are effective for capturing the nuances of speech[4].

Model Fusion and Decision Making**:** The integration of facial and speech data is achieved through a fusion strategy that combines the outputs of the individual models. This fusion can be performed at different levels, such as feature-level fusion, where features from both modalities are concatenated, or decision-level fusion, where the outputs of the models are combined to make a final decision. The fusion strategy enhances the system's ability to handle ambiguous or conflicting inputs, leading to more accurate emotion recognition.

Applications and future work:The dual-modality emotion recognition system has numerous applications, ranging from mental health monitoring, where it can help detect early signs of emotional distress, to enhancing virtual assistants and customer service platforms by providing emotionally intelligent responses. Future work could focus on expanding the system's capabilities to recognize a broader range of emotions, improving its real-time processing efficiency, and integrating it with wearable devices for continuous emotion monitoring.

3 METHODOLOGY

1. Facial Emotion Recognition

The facial emotion recognition component focuses on analyzing facial expressions from images or video frames. The steps involved are:

Image Preprocessing: Input images are preprocessed to normalize lighting conditions, scale, and orientation. This includes converting images to grayscale, resizing them to a standard dimension, and applying histogram equalization to improve contrast.

Feature Extraction: Convolutional Neural Networks (CNNs) are employed to extract relevant features from the facial images. CNNs are effective in capturing spatial hierarchies in images, making them suitable for detecting subtle changes in facial expressions. The CNN model is trained on labeled datasets where each image corresponds to a specific emotion.

Classification: The extracted features are passed through fully connected layers of the CNN, followed by a softmax activation function that outputs the probabilities of each emotion class. The emotion with the highest probability is selected as the predicted emotion for the input image.

2. Speech Emotion Recognition

The speech emotion recognition component processes audio signals to detect emotional states. The steps involved are:

Audio Preprocessing: Audio data is segmented into short frames to capture temporal variations in speech. Noise reduction techniques, such as spectral subtraction, are applied to minimize background noise.

Feature Extraction: Acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy, are extracted from the audio frames. MFCCs are widely used in speech processing due to their ability to represent the short-term power spectrum of sound.

Model Training: Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are used to model the sequential nature of speech data. LSTMs are effective in capturing long-term dependencies in time-series data, which is essential for understanding speech patterns over time.

Classification: The LSTM network outputs are fed into dense layers with a softmax function to classify the speech segments into different emotional categories[6].

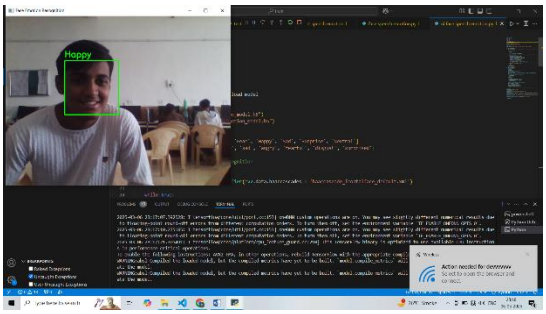
3. Data Fusion Strategy

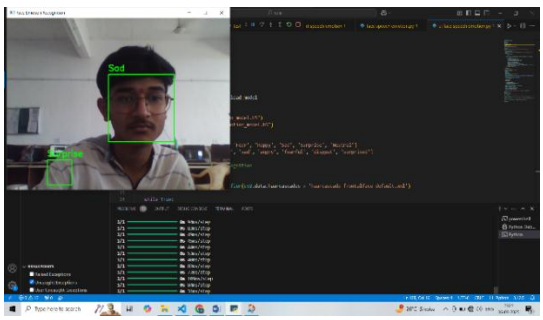
To improve the accuracy and robustness of emotion recognition, the outputs of the facial and speech emotion recognition components are combined using a fusion strategy. Two common fusion approaches are:

Feature-Level Fusion: The features extracted from both modalities are concatenated to form a single feature vector, which is then passed through a classifier to predict the emotion.

Decision-Level Fusion: The individual predictions from the facial and speech models are combined using techniques like weighted averaging or majority voting to determine the final emotion classification.

RESULT





4 TEST CASE

Test Case	Input Prompt	Expected Output	Result
Real-Time Face Detection	A person with a happy expression appears in front of the camera.	“emotion ": "happy”	PASS
Multiple Faces in Frame	Two or more people appear in front of the camera with different expressions.	"person_1":"happy""person_2":"sad"	PASS
No Face in Frame	The camera is pointed at an empty wall or object.	“error": "No face detected”	PASS
Changing Facial Expressions in Real-Time	Heavy server load A person changes their expression from neutral → surprise → sad within a few seconds.	"timestamp_1":"neutral""timestamp_2":"surprise""timestamp_3":"sad"	PASS

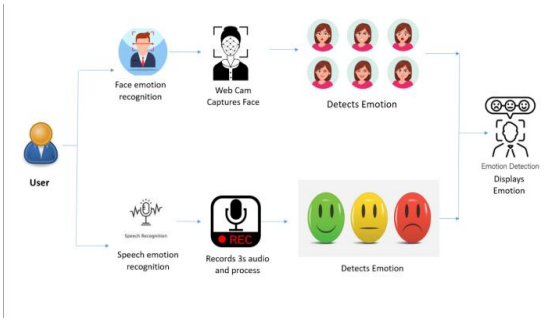
5 PROPOSED WORK

The Face and Speech Emotion Recognition project is designed to identify human emotions based on facial expressions and speech patterns. By integrating computer vision and speech processing techniques, this system can classify emotions such as happiness, sadness, anger, surprise, fear, and neutrality. The project leverages machine learning and deep learning models to analyze visual and audio inputs effectively. The first stage involves data collection from publicly available datasets such as FER-2013 and CK+ for facial emotions, and RAVDESS and TESS for speech emotions. In cases where additional data is required, custom datasets can be created by capturing images through a camera and recording voice samples. Once collected, the data undergoes preprocessing, where facial images are converted to grayscale, resized,

and normalized to enhance recognition accuracy. For speech data, features such as Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to represent the emotional aspects of voice signals.

For model development, deep learning approaches are employed. A Convolutional Neural Network (CNN) is used for facial emotion detection, as it excels in recognizing patterns in image data. Popular architectures like VGG-16 or ResNet are often used for improved feature extraction. Meanwhile, for speech emotion detection, an LSTM (Long Short-Term Memory) model processes sequential audio data. In some cases, a hybrid CNN-LSTM model is implemented to capture both spatial and temporal characteristics of speech. These models are trained using labeled datasets, and their performance is measured using accuracy, precision, recall, and F1-score metrics.

Once trained, the system is tested for real-time emotion detection. OpenCV is used to capture video input, detect faces, and apply the CNN model for facial emotion classification. For speech, a microphone records voice inputs, which are then processed using the trained LSTM model. The results from both facial and speech emotion recognition models can be combined for improved accuracy and a more comprehensive emotion assessment.To enhance user interaction, a Graphical User Interface (GUI) is developed using frameworks such as Flask or Streamlit. This interface allows users to visualize real-time emotion detection results. The system can also be integrated into applications like virtual assistants and customer service platforms to enhance user engagement. After rigorous testing, the project is deployed as a web-based application using Flask or FastAPI, or as a standalone desktop application. Cloud deployment and GPU acceleration can be considered for optimizing performance[6].



6 REQUIREMENT ANALYSIS

The Face and Speech Emotion Recognition system must fulfill both functional and non-functional requirements to ensure efficient performance. Functionally, the system should detect and classify emotions from facial expressions and speech inputs accurately. It must support real-time processing, allowing users to receive instant feedback on detected emotions. The facial emotion recognition component should be capable of identifying common emotions such as happiness, sadness, anger, surprise, fear, and neutrality by analyzing facial features. Similarly, the speech emotion recognition module should process voice signals to detect emotional states based on pitch, tone, and intensity variations. Additionally, the system should integrate both facial and speech emotion outputs to enhance recognition accuracy. A user-friendly interface should be developed to display the detected emotions, enabling seamless interaction. Furthermore, the system should allow data storage to maintain emotion records for analysis and model improvement[7,8].

In terms of non-functional requirements, the system must ensure high accuracy and precision in emotion classification to minimize errors. Performance optimization is crucial, particularly for real-time processing, where the model should analyze inputs quickly without significant delays. The system should be scalable, allowing it to handle increased data and user interactions efficiently. Usability is another critical factor; the interface must be simple and intuitive so users with minimal technical knowledge can operate it effectively. Security measures should also be implemented to protect sensitive facial and speech data, ensuring privacy compliance. Additionally, the system should be compatible with multiple platforms, such as Windows, macOS, and web-based applications, for broader accessibility. Overall, the system must be designed to deliver fast, accurate, and reliable emotion detection while maintaining user privacy and ease of use[9,10].

7. CONCLUSION

Face and speech emotion recognition is a rapidly evolving technology with significant implications across various industries. By leveraging artificial intelligence and deep learning, it enables systems to interpret human emotions accurately, leading to enhanced human-computer interaction, improved

mental health monitoring, and personalized user experiences. This technology bridges the gap between human emotions and digital responses, making interactions more intuitive and meaningful[11].

The integration of emotion recognition into healthcare, education, customer service, security, and entertainment has shown promising results in improving efficiency and engagement. In healthcare, it aids in early diagnosis of emotional distress, while in education, it helps assess student engagement. Businesses benefit from better customer interactions, and security agencies use it to detect potential threats. These advancements highlight the versatility and growing importance of emotion recognition systems in real-world applications. Despite its advantages, challenges such as accuracy, data privacy, and ethical concerns remain. Ensuring unbiased and ethical use of this technology is crucial for its responsible deployment. With continuous research and improvements, face and speech emotion recognition has the potential to revolutionize various sectors, making interactions more human-centered and enhancing the way we connect with technology.

8 FUTURE SCOPE

This research opens several promising directions for future development. Expanding to multilingual and culturally diverse datasets can enhance model accuracy and fairness. Optimizing the system for real-time performance on edge devices will increase its applicability in mobile and remote settings. Adding modalities like physiological signals or text can enrich emotion detection. Personalization based on individual emotional patterns may improve system relevance, especially in healthcare and education. Emphasis on ethical AI—ensuring fairness, transparency, and privacy—will be crucial for responsible deployment. Finally, integration with human-computer interfaces like virtual assistants and AR/VR can create more empathetic, adaptive user experiences.

REFERENCE

[1] Mobbs, Rebecca, Dimitrios Makris, and Vasileios Argyriou. "Emotion Recognition and Generation: A Comprehensive Review of Face, Speech, and Text Modalities." arXiv preprint arXiv:2502.06803 (2025).

[2] Rai, Alpana, Chetan Agrawal, and Divya Envey. "A Comprehensive Survey of Multimodal Emotion Recognition: Techniques, Applications, and Future Directions."

[3] Rajapakshe, Thejan, et al. "Parameterised Quantum Circuits for Novel Representation Learning in Speech Emotion Recognition." arXiv preprint arXiv:2501.12050 (2025).

[4] Fadhil, Muhammad Farhan, and Amalia Zahra. "The use of generative adversarial network as a domain adaptation method for cross-corpus speech emotion recognition." Bulletin of Electrical Engineering and Informatics 14.1 (2025): 297-306.

[5] Hasan, Rashedul, et al. "EmoFormer: A Text-Independent Speech Emotion Recognition using a Hybrid Transformer-CNN model." arXiv preprint arXiv:2501.12682 (2025).

[6] Kalaiyarasi, M., et al. "Student's Emotion Recognition using Multimodality and Deep Learning." ACM Transactions on Asian and Low-Resource Language Information Processing (2024).

[7] Sondawale, Shweta, et al. "Survey of Audio-Facial Emotion Decoder System."

[8] Lian, Hailun, et al. "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face." Entropy 25.10 (2023): 1440.

[9] Daneshfar, Fatemeh, and Mohammad Behdad Jamshidi. "An octonion-based nonlinear echo state network for speech emotion recognition in Metaverse." Neural Networks 163 (2023): 108-121

[10] Du, Yu, Rubén González Crespo, and Oscar Sanjuán Martínez. "Human emotion recognition for enhanced performance evaluation in e-learning." Progress in Artificial Intelligence 12.2 (2023): 199-211.

[11] Lim, Yixen, et al. "Emotion recognition by facial expression and voice: review and analysis." Journal of Informatics and Web Engineering 1.2 (2022): 45-54.