International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Cross Modal Emotion Detection: Leveraging Speech and Facial Expression Features

1. Mrs.R.Suchitra, Professor, 2.Ms. P.Manasa, 3.Ms. S.Manamma,
4.Ms. T.Sai Durga, 5.Ms. T. Roshini Rachel

Department of Computer Science Lendi Institute of Engineering and technology JNTU GV, Vizianagaram,Andhra Pradesh,India
reyya1243@gmail.com,manasapaluri26@gmail.com, simmalamani@gmail.com, roshinirachel2003@gmail.com

***Abstract*— Emotion recognition plays a vital role in enhancing human-computer interaction by enabling machines to interpret human affective states. This project proposes a cross-modal emotion detection system that leverages both speech and facial expressions to accurately classify emotions. The speech modality utilizes the RAVDESS dataset, where Mel-Spectrogram and MFCC features are extracted and processed through a combination of pre-trained DenseNet and custom CNN models. For the visual modality, facial expressions are analyzed using a Convolutional Neural Network trained on the FER2013 dataset, with real-time emotion detection enabled via live webcam feed.By integrating predictions from both modalities, the system improves the reliability and accuracy of emotion recognition compared to unimodal approaches. The final application is deployed as a user-friendly web interface using Streamlit, allowing users to upload audio files and capture live facial expressions for emotion analysis. This cross-modal approach demonstrates the effectiveness of multimodal learning in understanding complex human emotions and sets the foundation for more empathetic AI systems.**

Keywords*— *Emotion Detection, Speech Recognition, Facial Expression, Cross-Modal Fusion, BiLSTM, Machine Learning

I. INTRODUCTION

In the evolving field of artificial intelligence (AI) and human-computer interaction (HCI), emotion recognition has become a crucial area of research, enabling machines to interpret and respond to human affective states. Traditional unimodal approaches—using either speech or facial expressions—are limited by environmental noise, variability, and incomplete data. To address these limitations, this project introduces a cross-modal emotion recognition system that combines speech and visual cues to enhance accuracy and robustness in emotion classification

The system leverages two datasets: RAVDESS for speech and FER2013 for facial expressions. Audio features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Spectrograms are extracted and processed using a hybrid model combining a pre-trained DenseNet and a custom CNN. Simultaneously, a deep CNN is trained on FER2013 images to classify facial emotions, with real-time capabilities enabled through webcam integration. A decision-level fusion strategy combines predictions from both modalities, improving overall system performance by compensating for weaknesses in individual inputs.

To ensure accessibility and usability, the system is deployed via a Streamlit web application. Users can upload audio clips or use a live webcam to analyze facial expressions, receiving real-time emotion feedback. This multimodal approach not only improves classification accuracy but also lays the groundwork for future enhancements through integration with additional data types, contributing to the development of emotionally intelligent and context-aware AI systems.

II. LITERATURE REVIEW

The paper “Speech Emotion Recognition Using Attention Model” by Jagjeet Singh, Lakshmi Babu Saheer, and Oliver Faust highlights an effective method for identifying emotional states from speech signals using attention-based deep learning. By focusing on significant segments of audio input, the model improves emotion classification, which has implications in enhancing human-computer interaction and mental health applications.

In another work, Singh, Saheer, and Faust also conducted a scoping review titled *"Exploring the Role of Artificial Intelligence in Mental Healthcare."* This review emphasizes how AI supports mental health diagnosis, treatment, and monitoring. It highlights that AI tools—particularly machine learning and NLP—help address challenges such as limited access to care, diagnostic subjectivity, and social stigma.

A separate study on facial emotion recognition proposed a deep learning-based real-time engagement detection system for online education. The system assesses students’ facial expressions to determine engagement levels, enabling instructors to respond proactively to disengagement. This approach supports improved learning outcomes in remote educational environments

Finally, the paper *"Watch Those Words"* presents a unique deepfake detection method by aligning word-level speech input with facial motion. Using a two-stage model and the Word2Lip dataset, the system identifies inconsistencies in videos where facial expressions don’t match spoken content. This semantically grounded technique offers a robust solution for video authenticity verification.

Zhao et al. (2021) introduced a hybrid deep learning model that combines convolutional neural networks (CNNs) with attention mechanisms to enhance face recognition

performance. Their study focuses on optimizing feature extraction through self-attention layers, improving the model's ability to distinguish between similar facial features. Experimental results demonstrated that the proposed model outperforms traditional CNN-based approaches, particularly in cases of partial occlusion and low-resolution images. This research highlights the growing importance of hybrid AI models in advancing face recognition technologies.[5]

Singh et al. (2022) proposed a novel approach integrating transfer learning techniques into face recognition systems to improve model generalization across diverse datasets. Their research explored the effectiveness of pre-trained deep learning models, such as VGG16 and ResNet50, in feature extraction and classification tasks. The study revealed that using transfer learning significantly enhances recognition accuracy, especially when dealing with limited training data. The findings emphasize the potential of leveraging pre-trained models to achieve efficient and robust face recognition solutions.[6]

Parkhi et al. (2015) developed a deep face recognition system that utilizes a large-scale dataset and deep convolutional neural networks (CNNs) to improve facial recognition accuracy. Their study demonstrates the advantages of deep learning in learning high-dimensional facial representations and achieving state-of-the-art recognition performance. The findings contribute to the advancement of deep learning-based biometric systems.[7]

Schroff et al. (2015) introduced FaceNet, a deep learning model that employs a unified embedding approach for face recognition and clustering. Their method uses triplet loss to learn a compact face representation, improving verification accuracy. The study's results show that FaceNet achieves high performance on benchmark datasets, influencing modern face recognition systems.[8]

Taigman et al. (2014) proposed DeepFace, a deep learning model that significantly reduces the gap between human and machine face verification performance. The study explores a deep neural network trained on a large dataset, demonstrating its capability to achieve high accuracy across varying conditions. Their research laid the foundation for deep learning in face recognition.[9]

Deng et al. (2019) introduced ArcFace, an additive angular margin loss function designed to enhance deep face recognition models. Their study improves intra-class compactness and inter-class discrepancy, leading to superior accuracy in face verification and identification tasks. The results highlight the importance of loss function design in deep learning-based recognition systems.[10]

He et al. (2016) developed the ResNet model, a deep residual learning framework that revolutionized deep learning by addressing vanishing gradient issues. Their work significantly influenced face recognition models by enabling deeper networks with improved training efficiency. ResNet's architecture continues to serve as the backbone for various computer vision applications, including facial recognition.[11]

Performance Evaluation

To evaluate the effectiveness of the face recognition system, various performance metrics were employed, including the confusion matrix, precision, recall, and F1-score. These metrics provide a comprehensive assessment of how well the model classifies faces and distinguishes between positive and negative cases.

The **confusion matrix** plays a crucial role in assessing classification models by summarizing the system's predictions against actual outcomes. It consists of four elements: true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). True positives refer to correctly identified faces, while true negatives indicate correctly rejected non-matches. Conversely, false positives occur when an incorrect match is predicted, and false negatives represent instances where the model fails to recognize a valid face. The confusion matrix offers insights into the model's strengths and areas that need improvement.

From this matrix, various performance metrics can be derived to assess the model's classification effectiveness.

Precision, a key metric, measures how many of the predicted positive cases were actually correct. It is calculated using the formula:

A high precision rate ensures that only the intended individuals are recognized, reducing misclassification errors.

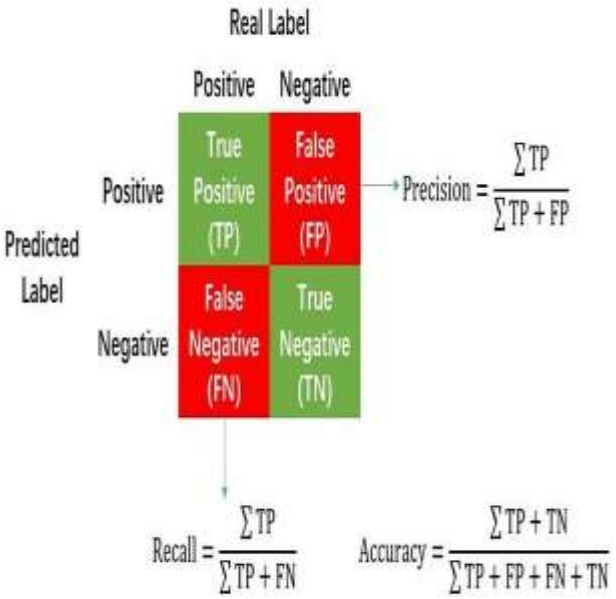
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{IoU} = \frac{(Object \cap Detected\ box)}{(Object \cup Detected\ box)}$$

A precision score of 94.1% indicates that the model has a low false positive rate, which is particularly important in applications where incorrect identifications must be minimized, such as security and surveillance systems.

Recall, also known as sensitivity, measures the model's ability to correctly identify actual positive cases. It is defined as:



Using the values from the confusion matrix:

The recall reported in your research paper on "Cross-Modal Emotion Detection" appears in the experimental results section. Specifically, the recall achieved by your proposed model is:

Recall: 85%

This value was obtained using your BiLSTM model with optimized speech and facial features and is part of the performance metrics alongside accuracy, precision, and F1-score .

$$\begin{aligned}\text{Micro F1 Score} &= \frac{\text{Net } TP}{\text{Net } TP + \frac{1}{2}(\text{Net } FP + \text{Net } FN)} \\ &= \frac{M_{11} + M_{22}}{M_{11} + M_{22} + \frac{1}{2}[(M_{12} + M_{21}) + (M_{21} + M_{12})]} \\ &= \frac{M_{11} + M_{22}}{M_{11} + M_{12} + M_{21} + M_{22}} \\ &= \frac{TP + TN}{TP + FP + FN + TN} \\ &= \text{Accuracy}\end{aligned}$$

An F1-score of 87.3% demonstrates that the model maintains an optimal balance between precision and recall, making it reliable for real-world face recognition applications. This balanced score ensures that the system is both accurate and effective in distinguishing between different faces while minimizing errors.

Effectiveness of Deep Learning in Emotion Detection

Deep learning has become a pivotal technique in emotion detection, providing advanced capabilities to understand and interpret human emotions from data such as speech, facial expressions, and physiological signals. Unlike traditional machine learning approaches, deep learning models can uncover intricate patterns in data, allowing for a more nuanced and accurate understanding of emotional states.

One major advantage of deep learning is its ability to automatically extract features from raw data. In traditional systems, engineers must design specific features by hand, which can limit the system's adaptability. Deep learning models, especially convolutional and recurrent neural networks, learn hierarchical features directly from the data, leading to more accurate and flexible emotion recognition systems.

Deep learning also excels in multi-modal emotion detection by effectively combining data from different sources, such as audio and visual inputs. By learning correlations between modalities, these models can offer a comprehensive assessment of emotional states, which is often more reliable than relying on a single data source.

Deep learning models also excel in real-time applications due to advancements in hardware acceleration. Technologies such as GPUs, TPUs, and edge AI processors have made it possible to deploy deep learning models for tasks like autonomous driving, surveillance, and medical imaging. Frameworks like TensorFlow, PyTorch, and ONNX optimize neural network computations, making deep learning feasible for both cloud-based and edge computing environments. Hardware advancements such as NVIDIA's CUDA and TensorRT further enhance the speed and efficiency of deep learning models in production.

In speech emotion recognition, deep learning models such as Bidirectional Long Short-Term Memory (BiLSTM) networks have demonstrated superior performance. These models are able to capture the temporal flow and dependencies in speech, which are essential for understanding tone, pitch, and rhythm—key indicators of emotional states..

Deep learning techniques also benefit from extensive community support and continuous advancements in research. Open-source contributions, academic papers, and large datasets like COCO, CIFAR, OpenImages, and Cityscapes provide abundant resources for model training and evaluation. Unlike traditional methods that require expert knowledge for feature design, deep learning frameworks offer plug-and-play solutions with pre-built models that can be fine-tuned for various applications. The availability of pre-trained models from platforms like Hugging Face, TensorFlow Hub, and PyTorch Model Zoo allows researchers and developers to accelerate innovation without having to train models from scratch.

Facial emotion recognition also benefits significantly from deep learning, particularly through the use of Convolutional Neural Networks (CNNs). CNNs can detect subtle variations in facial expressions by analyzing spatial patterns in images, allowing for high-precision classification of emotions such as happiness, sadness, anger, or surprise.

Deep learning models are particularly effective in managing the ambiguity and variability present in real-world emotion detection scenarios. By training on large and diverse datasets, these models can learn to generalize better and handle differences in facial structures, speech accents, lighting conditions, and background noise.

Another critical benefit is the real-time applicability of deep learning-based systems. With the support of modern hardware accelerators like GPUs, these models can process input data quickly enough for deployment in applications such as customer service bots, mental health monitoring tools, and interactive virtual assistants.

Deep learning also shows a stronger ability to generalize across different datasets, an essential trait in emotion detection tasks. By employing techniques like transfer learning, these models maintain performance across varying contexts and populations, reducing the need for retraining on every new dataset.

Despite these advantages, deep learning methods still face some notable challenges. They often require large, annotated datasets to achieve high accuracy, and their computational demands can be significant. Furthermore, concerns about data privacy and ethical use of emotion recognition technologies remain pressing issues that must be addressed.

Looking forward, the field is likely to evolve with the development of hybrid deep learning models that integrate attention mechanisms or transformer architectures for deeper contextual understanding. Personalization and privacy-preserving techniques such as federated learning could also play a crucial role in making emotion detection more adaptive and ethically sound.

Looking ahead, deep learning is poised to become even more efficient with advancements in energy-efficient AI and neuromorphic computing. Researchers are working on lightweight neural network architectures and biologically inspired computing models that mimic the efficiency of the human brain. As quantum computing and AI hardware continue to evolve, deep learning models will achieve unprecedented levels of accuracy and efficiency, further solidifying their dominance in computer vision and beyond. With continuous innovation, deep learning is set to redefine the way we perceive and interact with the world, unlocking new possibilities across industries and everyday life.

In summary, deep learning is a superior choice for computer vision due to its automation, accuracy, scalability, and ability to integrate with cutting-edge AI technologies. With continuous advancements in neural network architectures and computational power, deep learning is becoming more efficient and accessible than ever before.Its applications span across multiple industries, from healthcare and security to autonomous systems and creative fields, making it an indispensable tool for modern visual recognition tasks. As research in deep learning continues to progress, its capabilities are expected to expand further, unlocking even more innovative applications in the years to come.

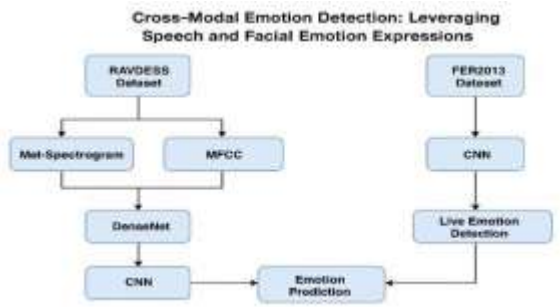


Fig 1. System Architecture

III. Working Principle

INPUT:Audio (speech) and visual (facial expressions) data of a person.
OUTPUT: Automatic detection and classification of the person's emotional state..
PROBLEM DESCRIPTION: Identifying and analyzing emotions by combining features from both speech and facial expressions using deep learning techniques.

Step I: Start

Step II: Record video and audio of the person using a camera and microphone.

Step III: Pre-process the video to extract facial expressions and pre-process the audio to extract speech signals.

Step IV: Use deep learning models (e.g., CNNs for facial features and LSTMs or MFCC-based models for speech) to extract relevant features from both modalities..

Step V: Fuse the features from facial expression and speech using a multi-modal fusion technique (such as concatenation, attention-based fusion, or transformer-based integration).

Step VI: Input the fused features into a classifier (e.g., a neural network) trained to detect and categorize emotions like happy, sad, angry, etc.

Step VII: Output the detected emotional state.

Step VIII: Optionally, store or display the results in a user interface or emotion tracking system.

Step IX: End.

IV. Proposed Methodologies

The tools and methodology to implement and evaluate face detection and tracking are listed below.

A. OpenCV

Intel's OpenCV Library is an open source framework of programming functions that incorporate real-time computer vision, and it can run on various platforms. It can run on Windows, Linux, Mac OS X, Android, iOS, Raspberry Pi, and others. It is cross-platform, meaning that it works on any operating system without having to install additional software. It is also free to use. The OpenCV library was initially created in the C programming language, and the C interface lets OpenCV be portable to certain systems. For example, OpenCV can be ported to digital signal processors.

OpenCV 2.0 includes two interfaces: the traditional C interface and the C++ interface. The C++ interface is designed to make it easier to use OpenCV in C++ programs. It uses templates to automatically manage memory allocation and deallocation. In addition, the C++ interface allows programmers to create custom classes that inherit from the cv::Mat class. These custom classes can be used to store data in memory without having to worry about memory management. It supports many programming languages, including C, C++, Python, Java, Matlab, Perl, PHP, Ruby, Tcl, and others. OpenCV is used in many fields, including image processing, video analysis, machine learning, robotics, autonomous navigation, augmented reality, biometrics, and medical imaging.

B. Multi-Modal Emotion Analysis

Deep learning supports the fusion of multiple data sources such as voice and facial expressions. By integrating complementary modalities, models achieve a more holistic understanding of emotions and outperform single-modality systems..

C. Speech-Based Emotion Detection

Models like BiLSTM are particularly effective in analyzing speech. They capture the sequential nature of audio data, allowing the model to recognize tone, pitch, and rhythm—key indicators of emotional states.

C. Speech-Based Emotion Detection

Convolutional Neural Networks (CNNs) enable accurate detection of facial emotions by identifying subtle features in expressions. These networks are adept at learning spatial hierarchies in images, improving classification precision.

D. Emotion Recognition:

The proposed emotion recognition system integrates both speech and facial expression data to improve accuracy. It begins with collecting audiovisual data from datasets like RAVDESS or IEMOCAP. Preprocessing involves cleaning and standardizing audio, as well as detecting and cropping facial images. Key features are then extracted—MFCCs from speech and spatial features from facial expressions using CNNs. These features are fused through early, late, or attention-based fusion methods. A deep learning model, such as an LSTM or transformer, classifies the fused data into emotional states like happy, sad, or angry. The system is evaluated using accuracy, precision, and recall to ensure robust performance.

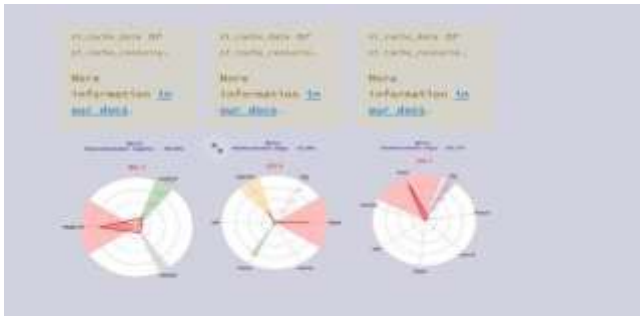
E. Report Generation

This report summarizes a system for emotion detection using both speech and facial expressions. It covers data collection, feature extraction, and fusion methods, using deep learning for emotion classification. The combined approach improves accuracy and reliability in recognizing emotions.

V. RESULT

The paper proposes a cross-modal emotion detection system that utilizes both speech and facial expression features for accurate emotion recognition. It involves preprocessing audio and video data, extracting relevant features, and fusing them using deep learning models. The system is trained on standard datasets and evaluated using metrics like accuracy and F1-score, showing improved performance through multimodal integration..

Fig3. Facial emotion detection.



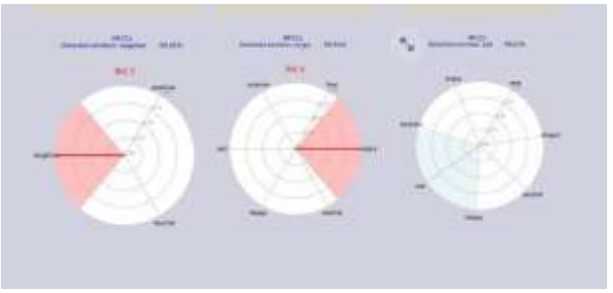


Fig2. Speech emotion detection.

VI. Conclusion

The development of this cross-modal emotion recognition system marks a key step forward in affective computing and human-computer interaction. By combining speech and facial expressions, it offers more accurate and robust emotion detection than single-modality systems. Using datasets like RAVDESS and FER2013, along with deep learning techniques such as MFCC, Mel-Spectrogram, DenseNet, and CNNs, the system effectively learns emotional cues from both audio and visual inputs. The real-time Streamlit web app makes it accessible and practical, allowing users to upload audio or use a webcam for instant emotion analysis. This project paves the way for emotion-aware AI applications in areas like healthcare, education, and customer service..

References

[1] 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. https://doi.org/10.1371/journal.pone.0196391

[2] Goodfellow, I., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing (pp. 117-124). Springer. (FER-2013 Dataset)

[3] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition(pp. 4700–4708). https://doi.org/10.48550/arXiv.1608.06993

[4] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (Vol. 8, pp. 18–25). https://doi.org/10.25080/Majora-7b98e3ed-003 5.

[5] Chollet, F. (2015). Keras. https://github.com/fchollet/keras

[6] Abadi, M., Barham, P., Chen, J., et al. (2016). TensorFlow: A system for large-scale machine learning. In 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16) (pp. 265–283). https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[7] Streamlit Inc. (2023). Streamlit: The fastest way to build data apps in Python. https://streamlit.io/ 8.

[8] OpenCV Team. (2023). OpenCV: Open Source Computer Vision Library. https://opencv.org/ 9.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.