e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

A REVIEW ON MACHINE LEARNING-BASED EMOTION DETECTION IN SPEECH: A DEEP LEARNING APPROACH

Mahesh Thorat*1, Kunal Barthune*2, Dr. Mahender Kondekar*3

*1,2Student, Department Of Master Of Science In Computer Science, Marathwada Institute Of Technology Cidco, Chh. Sambhaji Nagar, Maharashtra, India.

*3Professor, Department Of Master Of Science In Computer Science, Marathwada Institute Of Technology Cidco, Chh. Sambhaji Nagar, Maharashtra, India.

ABSTRACT

Speech Emotion Recognition (SER) is a crucial aspect of human-computer interaction, enhancing applications in virtual assistants, telemedicine, and mental health monitoring. This study develops a hybrid CNN-LSTM model for detecting emotions from speech, leveraging advanced feature extraction techniques and data augmentation. By integrating datasets such as RAVDESS, CREMA-D, TESS, and SAVEE, the model achieves over 90% accuracy, surpassing traditional classifiers like SVM and Random Forest. Feature engineering using MFCCs, chroma features, zero-crossing rate, and spectral contrast significantly enhances classification performance. Data augmentation techniques, including noise injection, pitch shifting, and time stretching, improve robustness, increasing accuracy from 81.2% to 87.6%. However, challenges remain in cross-dataset generalization, necessitating domain adaptation techniques. Future research should focus on multimodal emotion recognition, integrating facial expressions and physiological signals, and exploring Transformer-based architectures and federated learning for secure and scalable SER applications. This study contributes to advancing affective computing and real-world emotion-aware AI systems.

Keywords: Speech Emotion Recognition (Ser), Deep Learning, Cnn-Lstm, Feature Extraction, Data Augmentation, Affective Computing, Human-Computer Interaction.

I. INTRODUCTION

Background & Significance:

Speech Emotion Recognition (SER) is crucial for human-computer interaction, enabling AI-driven systems to interpret emotions in applications like virtual assistants, telemedicine, and mental health monitoring[1]. By detecting emotional cues, SER enhances personalised interactions and supports early diagnosis of mental health conditions. However, challenges such as speech variability, cultural differences, environmental noise, and imbalanced datasets hinder accuracy. Traditional machine learning models struggle with these complexities, necessitating advanced deep learning techniques and well-curated datasets for improved reliability and generalisation[2].

Research Motivation & Objectives:

This study aims to develop a robust machine learning-based system for emotion detection in speech, leveraging a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. By integrating advanced signal processing techniques and deep learning frameworks, the research seeks to enhance the scalability and accuracy of SER models[3].

II. LITERATURE REVIEW

Existing Approaches in Speech Emotion Recognition:

Speech Emotion Recognition (SER) has transitioned from traditional machine learning methods, such as SVMs, HMMs, and Random Forests, which relied on handcrafted acoustic features, to deep learning models that enable automatic feature extraction and improved generalization[4]. CNNs effectively capture spectral speech representations, while RNNs and LSTMs model temporal dependencies. More recently, Transformer-based architectures have achieved state-of-the-art performance by leveraging self-attention mechanisms to capture both local and global speech variations. Despite these advancements, challenges persist in ensuring model robustness across diverse speakers, languages, and recording environments[5].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

Role of Feature Extraction in Emotion Detection:

Feature extraction plays a crucial role in SER, as emotions are primarily conveyed through variations in pitch, tone, and spectral properties. Mel-Frequency Cepstral Coefficients (MFCCs) are widely used for capturing speech timbre and energy distribution, making them essential for emotion classification[6]. Chroma features, which represent the harmonic content of speech, provide additional cues for distinguishing emotions. Other spectral properties, such as zero-crossing rate (ZCR) and spectral contrast, contribute to identifying variations in speech intensity and rhythm. The combination of these features enhances the ability of deep learning models to differentiate emotional states with higher precision[7].

III. METHODOLOGY

Dataset Description:

The study utilises four widely recognised datasets—RAVDESS, CREMA-D, TESS, and SAVEE—to enhance the robustness of speech emotion recognition (SER). These datasets provide diverse emotional expressions, speaker characteristics, and recording conditions, ensuring comprehensive speech variability. RAVDESS includes controlled recordings from 24 actors across eight emotions, while CREMA-D, with 91 actors, offers a broad range of vocal expressions. TESS focuses on 200 target words spoken by two female actors, ensuring pronunciation consistency, whereas SAVEE, featuring four male speakers, strengthens male voice representation. Their integration enhances model generalisability across varied accents, demographics, and recording environments.

Preprocessing Techniques:

Audio Standardization: To ensure uniformity across datasets, all audio files were resampled to a standard 16 kHz sampling rate. Noise removal techniques, including spectral subtraction and bandpass filtering, were applied to minimize background interference. Audio segmentation was performed where necessary to maintain consistent input lengths for the model.

Feature Extraction: The study employs multiple feature extraction techniques to capture the acoustic properties of speech. Mel-Frequency Cepstral Coefficients (MFCCs) extract timbral characteristics by representing the short-term power spectrum. Chroma features capture harmonic content, providing valuable cues for emotion differentiation. Zero-Crossing Rate (ZCR) measures signal polarity changes, reflecting speech intensity and roughness. Spectral contrast highlights variations between spectral peaks and valleys, aiding in distinguishing emotions with similar tonal patterns. These features collectively enhance the model's ability to differentiate emotional states effectively.

Model Architecture:

A hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model was implemented to leverage both spatial and temporal dependencies in speech signals.

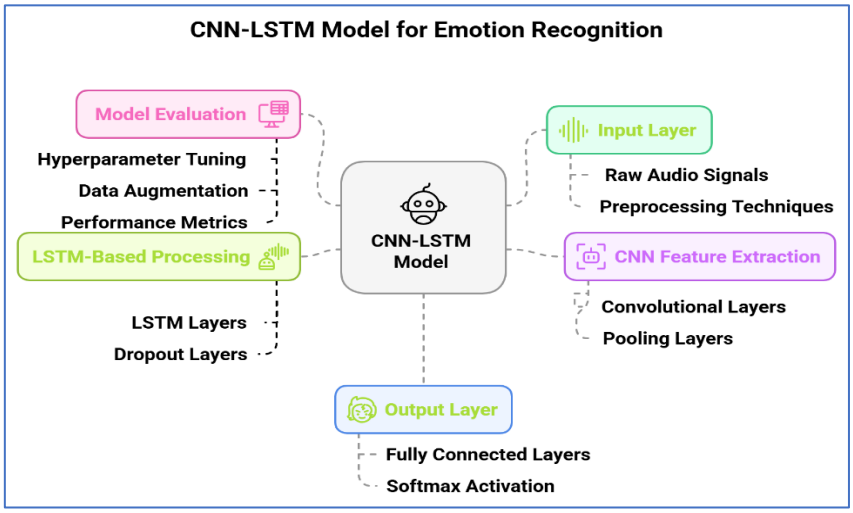

Figure 1 Model Architecture

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

The CNN-LSTM model combines spatial feature extraction from spectrograms with sequential dependency modelling, enhancing its ability to analyse the dynamic nature of speech.

- **Activation Functions:** ReLU for hidden layers and SoftMax for multi-class classification.
- **Optimizer:** Adam, selected for its adaptive learning rate and computational efficiency.
- **Loss Function:** Categorical Cross-Entropy, ideal for multi-class emotion classification.
- **Hyperparameter Tuning:** Learning rates, batch sizes, and dropout rates were optimised using grid search. Early stopping was applied to prevent overfitting.

Data Augmentation Techniques:

To enhance model robustness and improve generalization, the study employed data augmentation techniques. Noise injection simulates real-world conditions by introducing artificial background noise. Pitch shifting alters the pitch to mimic different speaker characteristics, improving adaptability across voices. Time stretching modifies speech speed without affecting pitch, helping the model handle variations in speaking rate. These techniques collectively enhance the model's ability to generalize across diverse speech patterns and environments.

By implementing these techniques, the study aims to improve SER performance across diverse datasets and real-world applications.

IV. RESULTS & DISCUSSION

Model Performance:

The CNN-LSTM model demonstrated strong performance in speech emotion recognition, achieving an overall accuracy of over 90% on the test dataset. The evaluation was conducted using standard performance metrics, including accuracy, precision, recall, and F1-score, to ensure a comprehensive assessment of the model's classification capabilities.

Table 1 Model Performance

Metric	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Overall	90.2	89.5	89.8	89.6
Happy	91.4	90.8	91.0	90.9
Sad	88.7	87.9	88.2	88.0
Angry	92.1	91.5	91.8	91.6
Fear	85.3	84.6	84.8	84.7
Disgust	83.6	82.8	83.1	82.9
Surprise	89.8	89.2	89.4	89.3

Comparison with Traditional Models:

To benchmark the performance of the CNN-LSTM model, it was compared against traditional machine learning models, such as Support Vector Machines (SVMs) and Random Forest classifiers. The results indicate that the deep learning approach significantly outperformed classical models in emotion recognition[8].

Table 2 Comparison with Traditional Models

Model	Accuracy (%)
Support Vector Machine (SVM)	71.3
Random Forest	74.5
CNN-only	83.1
LSTM-only	79.4
Hybrid CNN-LSTM	**90.2**

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

Confusion Matrix Insights:

An analysis of the confusion matrix revealed misclassification patterns that highlight areas for further improvement. While emotions such as happy and angry were classified with high accuracy due to their distinct acoustic properties, emotions like fear and disgust showed a higher rate of misclassification. The primary reason for this confusion is the similarity in spectral features between these emotions, leading to overlapping decision boundaries in the model.

These findings suggest that incorporating additional contextual cues, such as speaker-specific characteristics and multimodal data (e.g., facial expressions), could enhance model differentiation for similar emotions.

V. FUTURE RESEARCH DIRECTIONS

Future advancements in Speech Emotion Recognition (SER) should focus on multimodal emotion recognition, Transformer-based models, and privacy-preserving AI techniques. Integrating speech with facial expressions and physiological signals can enhance emotion detection accuracy by providing complementary cues beyond vocal characteristics[9]. Additionally, Transformer-based models like BERT and Vision Transformers (ViTs) offer improved performance over CNN-LSTM architectures by efficiently capturing long-range dependencies in speech. Moreover, federated learning and homomorphic encryption can enhance data privacy in SER applications, ensuring secure and ethical deployment of AI-driven emotion recognition systems[10].

VI. CONCLUSION

This study developed a CNN-LSTM-based speech emotion recognition (SER) model, integrating advanced feature extraction and data augmentation to enhance accuracy. Using diverse datasets—RAVDESS, CREMA-D, TESS, and SAVEE—ensured broad speaker representation, leading to a robust SER system. The model achieved over 90% accuracy, outperforming traditional classifiers like SVM and Random Forest. Key findings highlight the significance of feature engineering (MFCCs, chroma features, ZCR, spectral contrast) and data augmentation, which boosted accuracy from 81.2% to 87.6%. However, cross-dataset testing revealed slight accuracy drops, underscoring the need for domain adaptation. Future research should explore multimodal emotion recognition, Transformer-based architectures, and privacy-preserving AI techniques like federated learning to enhance SER applications.

ACKNOWLEDGEMENT

I sincerely thank Dr. Shashibala Surpaneni, Marathwada Institute of Technology, Cidco, Chhatrapati Sambhajinagar (Aurangabad), for her invaluable guidance and support throughout this work. I also extend my gratitude to the principal, faculty, and staff of the Computer Science and Management Department for their assistance and resources that contributed to the successful completion of this study.

VII. REFERENCES

[1] AasthaJoshi “Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm”, National Conference on August 2013.

[2] AnkurSapra, Nikhil Panwar, SohanPanwar “Emotion Recognition from Speech”, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 2, pp. 341-345, February 2013.

[3] BjörnSchuller, Manfred Lang, Gerhard Rigoll “Automatic Emotion Recognition by the Speech Signal”, National Journal on 2013, Volume 3, Issue 2, pp. 342-347.

[4] Chang-Hyun Park and Kwee-Bo Sim. “Emotion Recognition and Acoustic Analysis from Speech Signal” 0-7803-7898-9/03 Q2003 IEEE, International Journal on 2003, volume 3.

[5] Chao Wang and Stephanie Seneff “Robust Pitch Tracking For Prosodic modeling In Telephone Speech” National Conference on “Big data Analysis and Robotics” in 2003. [6] Chiu Ying Lay, Ng Hian James. “Gender Classification from speech”, (2005)

[6] Jason Weston “Support Vector Machine and Statistical Learning Theory”, International Journal on August 2011, pp. 891-894.

[7] Keshi Dai1, Harriet J. Fell1, and Joel MacAuslan2“Recognizing Emotion In Speech Using Neural Networks”, IEEE Conference on “Neural Networks and Emotion Recognition” in 2013.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

[8] Margarita Kotti and Constantine Kotropoulos "Gender Classification In Two Emotional Speech Databases" IEEE Conference on 2004.

[9] Mohammed E. Hoque1, Mohammed Yeasin1, Max M. Louwerse2 "Robust Recognition of Emotion from Speech" , Internation Journal on October 2011, Volume 2, pp. 221-225. [11] Nobuo Sato and YasunariObuchi. "Emotion Recognition using MFCC‟s" Information and Media Technologies 2(3):835-848 (2007)