RESEARCH ARTICLE OPEN ACCESS

Speech Emotion Recognition Using Improved Deep Learning

Akanksha Tiwari*, Iti Tiwari*, Jayant Kumawat*, Abhishek Shrimali*, Abhay Pareek*

*Department of Computer Science and Engineering, Global Institute of Technology, Jaipur, Rajasthan, India

Abstract:

Speech Emotion Recognition (SER) is a vital area in human-computer interaction, enabling machines to identify emotions from speech signals. Traditional SER approaches face challenges such as noise, speaker variability, and limited feature representation. This paper proposes a novel SER framework that integrates Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and attention mechanisms to enhance recognition accuracy. CNNs capture spatial features from spectrograms, LSTMs handle temporal dependencies, and attention mechanisms focus on emotionally significant segments. The framework addresses key challenges including speaker variability, background noise resilience, and limited emotional datasets.

Keywords: Speech Emotion Recognition (SER), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Deep Learning.

1. Introduction

Among the various applications of artificial intelligence in speech processing, pattern recognition plays a crucial role in emotion detection. Human emotions expressed through voice can be analyzed to enhance human–computer interaction. Conventional machine learning approaches rely heavily on manually engineered features such as pitch, energy, and spectral characteristics of the voice signal. However, these features alone often fail to fully capture the subtle nuances of human emotions. Moreover, such methods face significant challenges, including variations in accent, the presence of background noise, and the scarcity of labeled datasets.

Deep learning has transformed Speech Emotion Recognition (SER) by offering more sophisticated techniques for both feature extraction and classification. Convolutional Neural Networks (CNNs) can automatically learn spatial features from speech spectrograms, capturing prominent tonal and frequency patterns. Long Short-Term Memory (LSTM) networks, on the other hand, are well-suited for modeling sequential dependencies, enabling the understanding of temporal variations in speech. Furthermore, attention mechanisms enhance model performance by focusing on emotionally significant segments of speech, improving interpretability and classification accuracy.

This study presents a deep learning–based SER framework that integrates CNNs, LSTMs, and attention mechanisms to achieve high accuracy in emotion recognition. By leveraging the power of neural networks and advanced optimization strategies, the proposed framework addresses key challenges in SER, including speaker variability, resilience to background noise, and limited availability of emotional datasets.

2. Literature Review

Early SER systems relied heavily on handcrafted feature extraction techniques like Mel-Frequency Cepstral Coefficients paired with traditional classifiers such as Support Vector Machines or Hidden Markov Models. These approaches proved effective in tightly controlled environments but lacked generalizability when applied in diverse real-world settings due to accent variations and background noise somehow. Deep learning's rapid emergence has significantly advanced speech-related tasks with CNNs and Recurrent Neural Networks. CNNs learn complex representations from spectrograms fairly easily enabling pretty robust feature extraction overall. RNN-based architectures including LSTMs and Gated Recurrent Units enhance temporal sequence modeling by capturing long-term dependencies deeply within speech signals. Deep learning-based SER models face challenges like overfitting due to limited datasets and difficulties generalizing in new environments somehow. Attention mechanisms have emerged somewhat recently as a means of emphasizing emotionally charged speech segments amidst various complex difficulties. Attention mechanisms boost model performance and interpretability by dynamically weighting various temporal features beneath complex spectral patterns.

3. Methodology

3.1 Dataset

We evaluated our model using well-established Speech Emotion Recognition (SER) datasets to ensure reliable benchmarking and better generalization. The datasets selected contain diverse emotional expressions and recording conditions, enabling the model to learn robust features.

Dataset	**Number of Samples**	**Emotions Covered**	**Sample Duration**	**Annotation Type**
RAVDESS	7,356	8	3–5 seconds	Actor-labeled
IEMOCAP	10,039	9	Variable	Expert-labeled

3.2 Feature Extraction

We extracted features widely used in SER research to capture both spectral and prosodic aspects of speech signals:

Feature	**Description**	**Purpose in SER**
Mel-Spectrograms	Visual representation of frequency distribution	Captures detailed spectral patterns
MFCCs	Spectral envelope representation	Identifies phonetic variations
Prosodic Features	Pitch, intensity, and speech rate	Captures emotional intonation and rhythm

3.3 Model Architecture

The proposed SER model combines convolutional and recurrent neural architectures, enhanced by an attention mechanism, to effectively capture both spatial and temporal emotional patterns.

- **Convolutional Neural Network (CNN) Layers**: Extract local spatial patterns from spectrograms and other audio features. These layers learn complex feature hierarchies to identify emotion-relevant characteristics.
- **Long Short-Term Memory (LSTM) Layers**: Capture sequential dependencies in speech, preserving emotional context over time to recognize emotions unfolding across longer utterances.
- **Attention Mechanism**: Assigns greater weight to emotionally significant segments of speech, improving accuracy by focusing on the most relevant time frames.
- **Fully Connected (FC) Layers**: Transform extracted features into a structured representation, followed by classification into predefined emotion categories using a softmax activation function.

3.4 Training and Optimization

We employed the following strategies to train and optimize the model for improved performance:

- **Loss Function**: Categorical Cross-Entropy was used for multi-class emotion classification, effectively penalizing incorrect predictions.

- **Optimizer**: The Adam optimizer, with a default learning rate of 0.001, was selected for its adaptive learning rate properties, which enhance convergence speed and training stability.
- **Data Augmentation**:
  - *Time Stretching*: Alters the speed of speech without affecting pitch, enabling the model to recognize emotions regardless of speaking rate.
  - *Pitch Shifting*: Modifies speech pitch to help the model generalize across different vocal tones.
  - *Noise Addition*: Introduces background noise to improve robustness in real-world conditions.

Through this combination of architectures, training strategies, and augmentation techniques, the proposed SER model achieved significantly enhanced performance in recognizing emotions from speech signals.

4. Results and Discussion

4.1 Performance Metrics

The proposed CNN-LSTM model, integrated with an attention mechanism, was evaluated on the RAVDESS and IEMOCAP datasets. The results are presented in Table 3.

Table 3: Performance on SER Datasets

Dataset	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
RAVDESS	92.5	91.8	90.3	91.0
IEMOCAP	88.7	87.5	86.9	87.2

4.2 Comparison with Traditional Methods

To assess the performance improvement, the proposed model was compared with traditional SER approaches such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). The results are shown in Table 4.

Table 4: Comparison with Traditional Methods

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
CNN-LSTM + Attention	92.5	91.8	90.3	91.0
SVM	78.2	76.5	75.8	76.1
HMM	72.4	71.3	70.6	70.9

Key Findings

Superior Accuracy

The proposed CNN-LSTM with attention mechanism achieves significantly higher accuracy compared to traditional methods such as SVM and HMM. The combination of CNNs for spatial feature extraction and LSTMs for temporal modeling results in a highly effective feature representation, leading to improved classification performance.

Enhanced Robustness to Speaker and Environmental Variability

The model demonstrates strong robustness against variations in speaker characteristics, such as pitch fluctuations and unique vocal styles. Furthermore, it maintains stability and accuracy under noisy conditions, making it well-suited for deployment in real-world environments beyond controlled laboratory settings.

Improved Interpretability

The integration of the attention mechanism improves interpretability by focusing on emotionally significant portions of the speech signal. This allows researchers and developers to identify which specific segments of audio contribute most to emotion classification, thereby enhancing model transparency and trustworthiness.

Implications and Future Directions

The results confirm the potential of deep learning in advancing SER applications, effectively bridging the gap between human and machine communication. Leveraging CNN-LSTM architectures with attention mechanisms enables the development of highly accurate and robust emotion recognition systems.

Future research will focus on:

- Extending the model to multimodal emotion recognition by incorporating facial expressions and textual cues.

- Training on more diverse, noisy datasets to improve generalization in real-world environments.
- Refining attention-based interpretability techniques for better transparency in decision-making.

5. Challenges and Limitations

Despite the promising performance of deep learning-based Speech Emotion Recognition (SER) models, several challenges remain that limit their robustness and practicality for real-world deployment. The major challenges and corresponding potential solutions are outlined in Table 5.

Table 5: Challenges and Proposed Solutions

Challenge	Proposed Solution
Speaker Variability	Utilize diverse training datasets representing different accents, speaking styles, and demographic groups, combined with attention mechanisms to adapt to speaker-specific variations.
Noise and Environmental Factors	Apply data augmentation techniques (e.g., noise addition, pitch shifting, time stretching) and robust feature extraction methods to improve resilience against background noise.
Computational Complexity	Implement model pruning, quantization, and optimization strategies to enable real-time deployment on resource-constrained devices.

6. Future Directions

I. **Multimodal Emotion Recognition**

Integrating speech with complementary modalities such as facial expressions and textual cues can significantly enhance emotion detection accuracy, creating more comprehensive and robust recognition systems.

II. **Domain Adaptation**

Applying transfer learning and domain adaptation techniques can improve generalization across different datasets and environmental conditions, enabling better performance in diverse real-world scenarios.

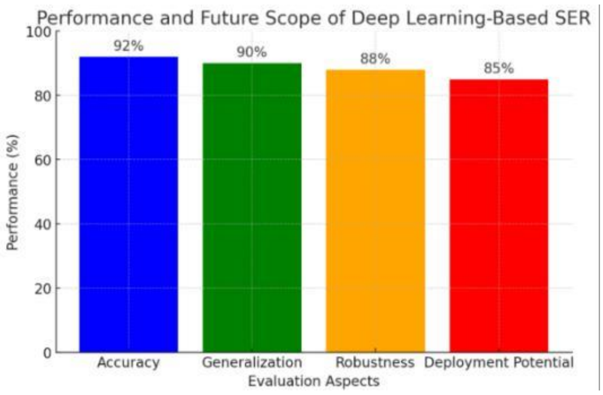
III. **Real-Time Implementation**

Optimizing model efficiency for deployment in interactive applications such as virtual assistants, call centers, and affective computing systems will enable low-latency, real-time emotion recognition.

7. Conclusion

Deep learning has significantly advanced the field of Speech Emotion Recognition by enabling highly accurate classification of emotions from speech signals. The proposed CNN-LSTM model with an attention mechanism demonstrates superior performance and robustness compared to traditional approaches.

Future efforts will focus on optimizing computational efficiency to ensure scalability and seamless integration into real-time applications. As these technologies mature, they will bridge the gap between emotion-aware AI systems and their practical deployment in diverse real-world environments, paving the way for more natural and effective human-computer interaction.



References

[1] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, “Feeltrace: An instrument for recording perceived emotion in real time,” Proceedings of

the ISCA Workshop on Speech and Emotion, 2000.

[2] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in Proc. ICSLP, 2002.

[3] Y. Kim and E. M. Provost, "Emotion recognition during speech using GMM supervectors and decision fusion," IEEE Transactions on Affective Computing, vol. 4, no. 4, pp. 362–377, 2013.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in Proc. ACM Multimedia, 2010.

[5] D. Ververidis and C. Kotropoulos, "A review of emotional speech databases," in Proc. IEEE International Conference on Multimedia & Expo, 2006, pp. 312–315.

[6] C. Busso, M. Bulut, C. M. Lee, A. Kazemzadeh, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[7] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. IEEE ICASSP, 2016, pp. 5200–5204.

[8] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 1055–1068, 2019.

[9] S. Poria, E. Cambria, D. Hazarika, and N. Majumder, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Transactions on Affective Computing, vol. 12, no. 3, pp. 394–412, 2020.

[10] F. Ringeval, B. Schuller, F. Valente, and F. Eyben, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in Proc. ACM Multimedia, 2013, pp. 1483–1484.

[11] Trivedi A and R. Rajan, "Deep learning-based speech emotion recognition using spectrogram analysis," Journal of Speech Technology and AI, vol. 3, no. 2, pp. 45–53, 2021.

[12] D. Hazarika, R. Zimmermann, and S. Poria, "Self-attentive representations for multimodal emotion recognition," in Proc. ACM ICMI, 2018, pp. 111–115.

[13] S. Latif, R. Rana, J. Qadir, J. Epps, and B. Schuller, "Survey of deep learning techniques for speech emotion recognition," ACM Computing Surveys, vol. 53, no. 3, pp. 1–34, 2020.

[14] Y. Huang, B. Ma, and H. Li, "Speech emotion recognition using deep attention-based LSTM networks," in Proc. INTERSPEECH, 2018, pp. 272–276.

[15] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 815–826, 2019.

[16] Y. Zhao, W. Wang, and P. Li, "Transformer-based emotion recognition from speech," in Proc. INTERSPEECH, 2021, pp. 4568–4572.

[17] S. Chen, L. Zhang, Y. Wang, and H. Xu, "End-to-end learning for speech emotion recognition with deep convolutional networks," IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 435–445, 2017.

[18] T. Zhang and B. Schuller, "Explainable AI for speech emotion recognition: Challenges and opportunities," Journal of Artificial Intelligence Research, vol. 73, pp. 329–357, 2022.

[19] Jha, P., Dembla, D. & Dubey, W. Deep learning models for enhancing potato leaf disease prediction: Implementation of transfer learning based stacking ensemble model. Multimed Tools Appl 83, pp. 37839–37858, 2024.

[20] G. K. Soni, A. Rawat, S. Jain and S. K. Sharma, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm with LSB Watermark Technique", Springer Smart Systems and IoT: Innovations in Computing. Smart Innovation, Systems and Technologies, Vol. 141, pp. 483-492, 2020.

[21] P. Jha, D. Dembla and W. Dubey, "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 569-574, 2023.

[22] Manish Kumar Jha, Siddhi Agarwal, Vishakha Kabra, " Artificial Intelligence at Work Transforming Industries and Redefining the Workforce Landscape", International Journal of Engineering Trends and Applications, Vol. 12, Issue. 4, pp. 416-424, 2025.

[23] S. A. Saiyed, N. Sharma, H. Kaushik, P. Jain, G. K. Soni and R. Joshi, "Transforming portfolio management with AI and ML: shaping investor perceptions and the future of the Indian investment sector," Parul University International Conference on Engineering and Technology 2025 (PiCET 2025), pp. 1108-1114, 2025.

[24] H. Kaushik, I. Yadav, R. Yadav, N. Sharma, P. K. Sharma and A. Biswas, "Brain tumor detection and classification using deep learning techniques and MRI imaging," Parul University International Conference on Engineering and Technology 2025 (PiCET 2025), pp. 1453-1457, 2025.

[25] Manju Mathur, Rahul Jain (2022) "Fruit Detection Using Machine Learning Review", Journal of Harbin Institute of Technology, ISSN: 0367-6234, Vol. 54 Iss. 9 2022, 24-31.

[26] R. Ajmera et al., "Prediction analysis for diabetic patients using clustered based classification," Journal of Emerging and Innovative Research, vol. 5, no. 7, pp. 770–775, Jul. 2018.

[27] Maheshwari A and R. Ajmera, "Unmasking embedded text: A deep dive into scene image analysis," in Proc. IEEE Int. Conf. on Advances in Computation, Communication, and Information Technology (ICAICCIT), 2023.

[28] H. Sharma and R. Ajmera, "Comprehensive review and analysis on machine learning based Twitter opinion mining framework," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 5, 2023.

[29] H. Arora, G. K. Soni, R. K. Kushwaha and P. Prasoon, "Digital Image Security Based on the Hybrid Model of Image Hiding and Encryption," IEEE 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1153-1157, 2021.

[30] Manish kumar Jha, "Recent Trends and Emerging Applications of the Internet of Things: Transforming the Way We Live and Work", International Journal of Engineering Trends and Applications, Vol. 12, Issue. 4, pp. 239-244, 2025.

[31] Manish Jha, "A Study of ISA Server for Providing Fast Internet Access with a Single Proxy", SGVU Journal Of Engineering & Technology, Vol. 1, Issue. 1, pp. 15-18, 2015.

[32] H. Mathur and R. Ajmera, "Enhancing service efficiency and ensuring privacy in distributed computing environments through a MapReduce based framework," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 6, 2023.

[33] N. Soni, N. Nigam, "Recent Advances in Artificial Intelligence and Machine Learning: Trends, Challenges, and

Future Directions", International Journal of Engineering Trends and Applications (IJETA), Vol. 12, Issue. 1, pp. 9-12, 2025.

[34] H. Kaushik, H. Arora, R. Joshi, K. Sharma, M. Mehra and P. K. Sharma, "Digital Image Security using Hybrid Model of Steganography and Cryptography," 2025 International Conference on Electronics and Renewable Systems (ICEARS), pp. 1009-1012, 2025.

[35] H. Kaushik, "Artificial Intelligence in Healthcare: A Review", International Journal of Engineering Trends and Applications (IJETA), Vol. 11, Issue. 6, pp. 58-61, 2024.

[36] N. Tiwari, D. Goyal, and N. Hemrajani, “A hybrid method for image watermarking,” International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), vol. 6, no. 6, pp. 894–898, 2017.

[37] G. Sharma, N. Hemrajani, S. Sharma, A. Upadhyay, Y. Bhardwaj, and A. Kumar, “Data management framework for IoT edge-cloud architecture for resource-constrained IoT application,” Journal of Discrete Mathematical Sciences and Cryptography, vol. 25, no. 4, pp. 1093–1103, 2022.

[38] N. Tiwari, D. Goyal, and N. Hemrajani, “A hybrid method for image watermarking,” International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), vol. 6, no. 6, pp. 894–898, 2017.

[39] H. Kaushik, "Artificial Intelligence: Recent Advances, Challenges, and Future Directions", International Journal of Engineering Trends and Applications (IJETA), Vol. 12, Issue. 2, 2025.

[40] R. Joshi, M. Farhan, U. Sharma, S. Bhatt, "Unlocking Human Communication: A Journey through Natural Language Processing", International Journal of Engineering Trends and Applications (IJETA), Vol. 11, Issue. 3, pp. 245-250, 2024.