IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Evaluating The CNN-LSTM Hybrid Architecture For Robust Speech Emotion Recognition

1st Author **: Miss.Borhade Dnyaneshwari Ravindra**,2nd Author:**Miss.Bhujbal Rutuja Santosh**

JCEI's Jaihind Institute of Management and Research (MCA)

3rd Author **: Prof.Dnyaneshwar Balu Lokhande(Research Guide),**4th Author**:Prof.Shubhangi Pratik Bombale(Research Guide)**

Abstract: Human speech is a fundamental mode of communication, rich not only in linguistic data but also in emotional cues. **Speech Emotion Recognition (SER)** aims to automatically decode emotional states by analyzing vocal characteristics such as spectral shape, rhythm, intensity, and pitch. With the capabilities of modern Artificial Intelligence (AI), deep learning approaches provide superior methods for extracting complex emotional patterns compared to earlier machine learning techniques. This study introduces a hybrid **Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) architecture** designed for robust emotion classification. The proposed methodology involves extensive preprocessing and the extraction of multiple features, including Mel-Frequency Cepstral Coefficients (MFCC), Chroma, and Zero Crossing Rate (ZCR). Trained on a fused dataset including benchmark resources like RAVDESS, TESS, and CREMA-D , the CNN-LSTM model demonstrated an overall accuracy of **90%+**. This result confirms the model's strong generalization and significant performance improvement over classical methods, offer ing a valuable contribution to emotion-aware AI systems for applications in human–computer interaction (HCI) and mental health analysis.

Index Terms – HCI, Convolutional Neural Network, Acoustic Variability,Hidden Markov Models

I. Introduction

Emotion recognition is vital across multiple input modalities, but speech is arguably the most non-intrusive and accessible carrier of emotional information. SER systems are effective even in scenarios limited to audio, such as call centers or virtual learning environments. The field has moved beyond relying on traditional Machine Learning (ML) models like SVM and HMM, which depended on manually crafted features, to advanced Deep Learning (DL) architectures. DL models, including CNNs and LSTMs, have revolutionized SER by automatically extracting hierarchical features and effectively modeling long-term temporal dependencies in unstructured audio.

1.2 Problem Statement

Despite these advancements, SER technology faces persistent challenges:

- **Acoustic Variability:** Emotional expressions vary significantly in modulation, duration, and intensity, making consistent classification difficult[13].
- **Similar Emotional Tones:** Emotions can be acoustically ambiguous; for example, "Calm" and "Sad" often share similar low-energy profiles, while "Angry" and "Fear" may overlap in high-energy patterns14.

- **Model Generalization:** Models trained on acted, high-quality datasets (like RAVDESS) frequently perform poorly when deployed in real-world environments contaminated by noise (echoes, traffic, background voices)[15151515].

This study aims to overcome these limitations by using a CNN-LSTM hybrid architecture trained on a fusion of multiple datasets and a rich set of high-quality features[16].

1.3 Hypothesis

This research is designed to test the following premise:

- **Alternative Hypothesis (H_1):** A CNN-LSTM-based hybrid deep learning model significantly improves accuracy, generalization, and robustness in emotion recognition compared to traditional machine learning and non-hybrid deep learning models.

2. Literature Review

SER research can be segmented into three distinct eras, each marked by new methodologies and models.

2.1 Early Machine Learning Approaches (Pre-2010)

Early SER systems relied on hand-engineered acoustic features such as pitch, MFCC, and Linear Predictive Coding (LPC).

- **Hidden Markov Models (HMM):** These were effective for modeling speech dynamics as sequences of acoustic states , but their performance was constrained by the independence assumption.
- **Support Vector Machines (SVM):** Strong early classifiers for high-dimensional features, SVMs were limited by their dependence on complex feature engineering and poor scalability with large datasets.

Algorithm	**Strengths**	**Limitations**
HMM	Good temporal modeling [23]	Independent-state assumption, low accuracy [24]
SVM	Strong classifier for small datasets [25]	Requires complex handcrafted features [26]

[Table 1 – Summary of Early Machine Learning Approaches]

2.2 Deep Learning Transition (2010–2018)

The introduction of deep neural networks allowed for automatic feature learning:

- **Convolutional Neural Networks (CNN):** Excels at extracting spatial features from 2D spectral representations (like spectrograms or MFCC matrices) but lacks the ability to capture temporal sequences.
- **Long Short-Term Memory (LSTM):** A refinement of Recurrent Neural Networks (RNN), LSTMs use memory gates to overcome vanishing gradients, allowing them to effectively model long-range temporal dependencies in the speech signal.

Model	Contribution	Drawbacks
CNN	Learns spatial-spectral emotion patterns [29]	No temporal sequence modeling [30]

Model	Contribution	Drawbacks
LSTM	Captures long-range temporal cues [31]	Computationally expensive for deployment [32]

[Table 2 – Key Contributions of Deep Learning Models]

2.3 Modern SER Approaches (2019–2024)

Current research favors hybrid and attention-based models:

- **CNN-LSTM Hybrid Models:** These systems utilize the CNN for initial feature extraction and the LSTM for subsequent sequence modeling. Studies indicate they achieve a **10–15% accuracy improvement** over standalone DL models.
- **Transformer-based Models:** Architectures like Wav2Vec2.0 and AST leverage self-attention mechanisms to learn contextual relationships without recurrence, achieving state-of-the-art performance.

3. Research Methodology

The research employed a descriptive and experimental design.

3.1 Data Acquisition and Preprocessing

The study utilized a **Dataset Fusion** strategy to enhance generalization and reduce speaker-dependent bias. The combined datasets include:

- **RAVDESS:** Acted audio-visual data (primary training data).
- **TESS:** Studio-recorded data (cross-validation).
- **CREMA-D:** Multilingual, realistic data (real-world generalization test).

A standard **Preprocessing Pipeline** was applied to the raw audio signals:

- **Noise Reduction:** To mitigate background noise and echoes.
- **Normalization:** To ensure uniform energy levels across all samples.
- **Silence Removal (VAD):** Using Voice Activity Detection (VAD) to reduce data redundancy and computational load.

3.2 Feature Extraction

The model was trained on a robust, concatenated feature vector:

- **MFCC (Mel-Frequency Cepstral Coefficients):** Crucial for capturing the spectral envelope and vocal tract shape.
- **Chroma:** Represents energy distribution across pitch classes, relevant for the "musicality" of the speech.
- **Zero Crossing Rate (ZCR):** Measures signal frequency, often related to unvoiced speech.
- **Spectral Centroid:** An indicator of the spectral shape and "brightness" of the sound.

3.3 Model Development: CNN-LSTM Hybrid

The hybrid model is structured as follows: | Component | Function | Benefit in SER | | :--- | :--- | :--- | | **CNN** | Extracts spatial/spectral features from 2D input (e.g., MFCC matrix). | Automatically learns hierarchical emotional patterns. | | **LSTM** | Processes the sequence of features output by the CNN. | Captures long-term temporal dependencies (e.g., rhythm and intensity changes). | | **Hybrid** | CNN acts as a feature extractor, feeding the sequence into the LSTM. | Provides strong temporal-spatial learning, resulting in significant accuracy improvement. | [Table 7 - Model Development: CNN-LSTM Hybrid Architecture]

4. Results and Discussion

The model was tested using an **80% Training / 20% Testing** data split.

4.1 Overall Performance and Comparison

The proposed CNN-LSTM model achieved superior overall accuracy compared to both traditional and standalone deep learning models:

Model	**Accuracy (%)**	**Performance Category**
SVM	72% [57]	Early Classical ML [58]
CNN	84% [59]	Deep Learning Era (Spatial Only) [60]
LSTM	87% [61]	Deep Learning Era (Temporal Only) [62]
CNN-LSTM	**90%** [63]	**Modern Hybrid Approach** [64]

Metric	**Score**	**Interpretation**
Accuracy	**90.0%** 66	90 out of 100 emotional samples classified correctly[67].
Precision (Macro Avg)	89.5% [68]	Low rate of false positives across all emotion classes[69].
Recall (Macro Avg)	89.8% [70]	High rate of correctly identifying true positive emotional instances[71].
F1-Score (Macro Avg)	89.6% [72]	Robust balance between precision and recall[73].

[Table 4 – Metric evaluation matric]

4.3 Discussion of Findings

1. **Hybrid Effectiveness:** The 90% accuracy confirms the core finding that the hybrid architecture successfully integrates the strengths of the CNN (spectral pattern recognition) and the LSTM (temporal sequence modeling)[74].

2. **Addressing Misclassification:** The inclusion of diverse features such as Chroma and Spectral Centroid alongside MFCC provided the richer discriminatory information necessary to reduce confusion between similar-sounding emotions (e.g., Sad vs. Calm)[75].

3. **Computational Balance:** While state-of-the-art Transformer models achieved a theoretical 94% accuracy, the CNN-LSTM model's 90% performance provides an **optimal trade-off for lightweight deployment**[767676]. This makes the model highly suitable for industrial applications requiring low computational cost on edge devices[77].

4. **Generalization:** Training on the diverse, fused dataset, particularly the inclusion of the realistic CREMA-D data, significantly enhanced the model's robustness against speaker-dependent variations and prepared it for real-world acoustic variability[78].

5. Conclusion and Future Scope

The experimental results fully validate the Alternative Hypothesis (H_1)[79]. The **CNN-LSTM hybrid model is confirmed as a robust and high-performing architecture** for Speech Emotion Recognition, demonstrating high accuracy and improved generalization across diverse datasets[80].

Future research should focus on the following:

- **Integration of Attention Mechanisms:** Introducing an Attention Mechanism layer into the CNN- LSTM is expected to selectively weight the most emotionally significant speech segments, potentially pushing accuracy to 92% or 94%[81].
- **Multimodal Fusion:** Extending the framework to fuse audio features with visual (facial expressions) and textual data to improve robustness and accuracy in ambiguous real-world scenarios[82].
- **Real-Time Optimization:** Applying model compression techniques, such as **quantization and pruning**, to further reduce the model's inference latency, thereby enabling true real-time, lightweight deployment on mobile or IoT devices

6. References

Ref. No.	Citation (IEEE Style)
[1]	L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989. (Foundation for HMM and GMM)
[2]	J. A. Schuller et al., "Recognizing speech emotion with long short-term memory (LSTM) recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 493-497. (Foundation for LSTM in SER)
[3]	H. S. K. Chen et al., "Deep learning for acoustic feature extraction in speech recognition: a comparison with MFCC," in *Proc. Interspeech*, 2013, pp. 202-205. (Foundation for MFCC features)

[4]	A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645-6649. (Foundation for RNNs)
[5]	T. M. K. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436- 444, May 2015. (General Deep Learning foundation)