Real-Time Speech Emotion Recognition with a CNN-BiLSTM-Attention Deep Learning Model

Hmad Zennou
TIAD, Sultan Moulay Slimane University, Beni Mellal, Morocco
hmannou2007@gmail.com (corresponding author)

Raja Ouadad
LIMATI, Sultan Moulay Slimane University, Beni Mellal, Morocco
ouadadraja2@gmail.com

Mohamed Ouhda
TIAD, Sultan Moulay Slimane University, Beni Mellal, Morocco
med.ouhda@gmail.com

Mohamed Baslam
TIAD, Sultan Moulay Slimane University, Beni Mellal, Morocco
m.baslam@usms.ma

Received: 12 January 2026 | Revised: 9 February 2026, 28 February 2026, 16 March 2026, 27 March 2026, and 31 March 2026 | Accepted: 1 April 2026

 DOI: https://doi.org/10.48084/etasr.17495

ABSTRACT

Speech Emotion Recognition (SER) aims to automatically identify human emotions from audio signals by leveraging advanced artificial intelligence techniques. Speech contains multiple layers of information, such as prosodic variation, voice quality, and spectral patterns, captured through continuous and spectral features. Selecting the most informative features is crucial to accurately modeling emotional expression. Many SER systems rely primarily on spectral features, such as MFCCs; however, this study combines both MFCC and RMSE features to construct a richer emotional representation. A hybrid CNN-BiLSTM-Attention architecture is proposed, which integrates convolutional layers for extracting local spectral patterns, a bidirectional LSTM for capturing long-range temporal dependencies, and a soft attention mechanism that emphasizes the most relevant segments of speech. Experimental evaluation on the RAVTESS dataset demonstrates that the proposed model achieved 98.10% accuracy, 97.95% precision, 98.02% recall, and a 97.98% F1-score, outperforming baseline CNN-LSTM models. Although the model is lightweight and designed for real-time suitability, explicit inference latency and throughput measurements are reserved for future work. These results confirm that integrating attention improves recognition of emotionally salient cues, yielding a robust and compact framework suitable for practical SER applications.

Keywords-speech emotion recognition; attention; CNN; BiLSTM; deep learning; human emotions

I. INTRODUCTION

Speech is one of the richest and most natural forms of human communication, conveying not only linguistic information but also emotional and psychological cues [1, 2]. Automatically recognizing emotions from speech, commonly referred to as Speech Emotion Recognition (SER), has become an important research area with applications in human-computer interaction, healthcare monitoring, call-center automation, intelligent tutoring systems, and affect-aware robotics [3-5]. Despite notable progress in recent years, SER remains a challenging task due to the inherent complexity of emotional expression, speaker variability, background noise, and the nonlinear and time-varying nature of speech signals [6].

To interpret emotional cues, SER systems rely on acoustic features extracted from speech signals. These features generally include prosodic attributes (e.g., pitch, energy, and duration), voice-quality measures (e.g., jitter, shimmer, and harmonic-to-noise ratio), and spectral representations such as Mel-Frequency Cepstral Coefficients (MFCCs). Among these, MFCCs have been widely adopted due to their ability to capture frequency-domain characteristics associated with emotional states [7]. However, relying solely on spectral features may overlook important temporal and energy-related variations that are crucial for distinguishing subtle emotional

expressions. Recent studies have emphasized the importance of combining multiple complementary feature types to obtain more discriminative emotional representations [8].

Early SER systems primarily employed handcrafted acoustic features coupled with traditional machine learning classifiers such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM). Although these approaches achieved reasonable performance, their limited capacity to model long-range temporal dependencies and generalize across speakers and recording conditions constrained their effectiveness [9]. In [10], a SER method incorporated relative difficulty and labeling reliability to improve generalization on unseen corpora. In [11], self-supervised speech representations were combined with spectral features using a mixture-of-experts framework, addressing domain-shift challenges in SER.

With the advancement of deep learning, neural network-based approaches have become dominant in SER research. Convolutional Neural Networks (CNNs), in particular, have demonstrated strong capability in learning discriminative spatial patterns from speech representations such as spectrograms and mel-spectrograms [12]. Several studies have explored CNN-based SER models on the RAVDESS dataset. For instance, in [13], a multi-task gated residual network used spectrograms extracted from speech and songs, achieving an accuracy of 65.97%. Similarly, in [14], a VGG-16 architecture was trained on speech spectrograms, reporting an accuracy of 71%. To further enhance CNN-based performance, in [15], multiple transformed audio features, including MFCCs, chromagrams, mel-spectrograms, spectral contrast, and tonnetz representations, were integrated, achieving an accuracy of 71.61%. However, these models primarily focused on spatial feature learning and largely neglected the temporal dependencies inherent in emotional speech.

To address this limitation, hybrid architectures combining CNNs with recurrent neural networks, such as Long Short-Term Memory (LSTM) networks, have been proposed. These models jointly capture local spectral patterns and temporal dynamics of speech signals. For example, in [16], a CNN-LSTM framework reported an accuracy of 64%, which was further improved to 66.5% and 70.8% by incorporating attention mechanisms and semantic embeddings, respectively. Bidirectional LSTMs (BiLSTMs) have also been employed to exploit both past and future contextual information, leading to improved emotion recognition performance [17, 18].

More recently, attention mechanisms, Transformer-based architectures, and self-supervised learning approaches have gained increasing attention in SER research. Attention mechanisms enable models to focus on emotionally salient segments of speech, improving robustness and interpretability [19, 20]. Transformer-based models have demonstrated effectiveness in modeling long-range dependencies in spectrogram representations [14]. In parallel, self-supervised speech representation models such as wav2vec 2.0 [21] and HuBERT [22] have shown strong capability in learning rich contextual embeddings from raw audio, significantly improving SER performance when used as pretrained feature extractors [23, 24]. Nevertheless, these approaches often require large-scale pretraining and substantial computational resources, which may limit their practicality for real-time or resource-constrained applications.

Hybrid architectures combining CNNs, LSTMs, Transformers, and attention mechanisms have also been explored. For instance, in [25], a parallel CNN-Transformer architecture used MFCC features, while in [26], a time-distributed CNN-Transformer model achieved an accuracy of 82.72% on RAVDESS. Other studies have demonstrated that combining advanced data augmentation techniques with attention-enhanced architectures can further improve SER performance [27]. Although these models achieve high accuracy, they often involve many trainable parameters, increasing computational complexity.

Despite these advances, several limitations remain. Many existing SER systems rely predominantly on spectral features and overlook the complementary role of continuous or energy-based features. Moreover, relatively few studies fully exploit the synergistic benefits of combining CNN, BiLSTM, and attention mechanisms within a lightweight and unified framework that balances performance and computational efficiency. This work addresses these limitations by proposing a hybrid CNN-BiLSTM-Attention architecture that integrates both spectral and continuous features for robust SER. The CNN component captures local spectral-temporal patterns, the BiLSTM models long-range emotional dynamics in both temporal directions, and the attention mechanism selectively emphasizes emotionally informative speech segments. The proposed approach is evaluated on the RAVDESS dataset and the combined RavTess dataset, demonstrating strong generalization across both single-dataset and cross-dataset settings. The main contributions of this work are summarized as follows:

- Presents a lightweight CNN-BiLSTM-Attention architecture that effectively captures spectral, temporal, and contextual emotional cues while maintaining low computational complexity suitable for real-time SER.
- Introduces a compact multi-feature fusion strategy using MFCC and RMSE features, demonstrating that a carefully selected feature set can outperform richer but redundant representations.
- Conducts extensive experiments on both single-dataset (RAVDESS) and cross-dataset (RavTess) scenarios, showing strong robustness and generalization across heterogeneous emotional speech data.
- Provides a comprehensive comparison with state-of-the-art deep learning models in terms of both recognition performance and model size.

II. MATERIALS AND METHODS

This section outlines the methodological framework adopted to develop the proposed SER system based on a CNN-BiLSTM-Attention architecture. Figure 1 illustrates the complete workflow. Emotional speech samples were first collected from the RAVDESS dataset, which served as the primary benchmark for experimentation. The audio files then

underwent standard preprocessing procedures, including noise reduction, normalization, framing, and spectrogram transformation, to ensure consistency and suitability for feature extraction.

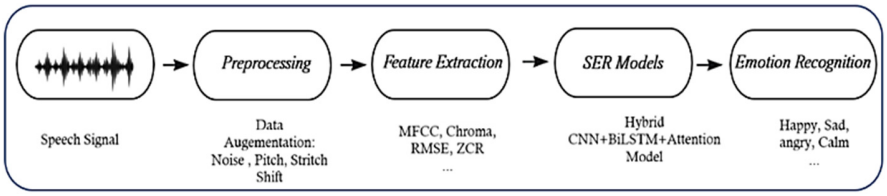

Fig. 1. Proposed method for SER.

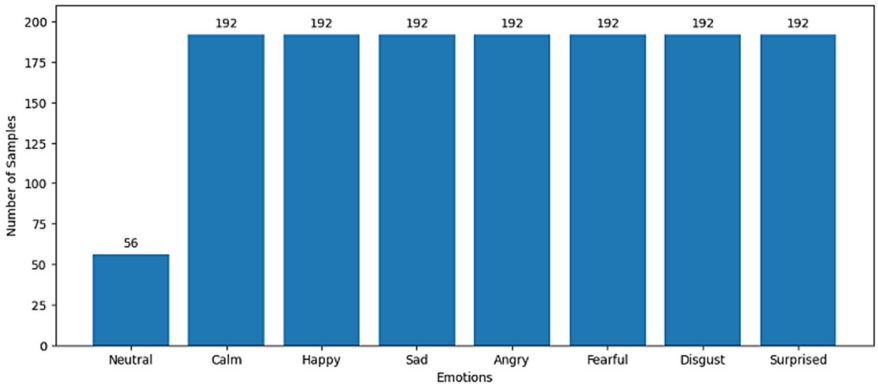

Fig. 2. Count plot graph of emotions in the RAVDESS dataset.

From the preprocessed signals, a diverse set of acoustic features is extracted to capture both spectral and temporal features relevant to emotional expression. These features are subsequently integrated through a multi-feature fusion mechanism, allowing the model to leverage complementary information and enhance representation quality. The fused feature set is then fed into the proposed deep learning architecture. Convolutional layers (CNN) are used to learn local spatial patterns from the feature maps. These representations are passed to BiLSTM layers to capture long-range temporal dependencies in both forward and backward directions. An attention mechanism is finally applied to emphasize the most informative segments of the speech signal and reduce the impact of irrelevant or noisy components. This combination enables the model to effectively identify subtle emotion-related cues.

A. Dataset

1) RAVDESS

This study uses the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [28], which includes recordings from 24 actors (12 male, 12 female) expressing eight emotions: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. Only the speech recordings were used. The dataset is available in three formats—audio-only, audio-video, and video-only—but all experiments were conducted on 16-bit, 48 kHz WAV files to ensure consistent acoustic analysis. Each sample lasts around 4 s with brief silence padding. For the emotions happy, sad, and angry, 192 speech files are available for each, totaling 576 samples. Features were extracted from every audio file to build the training set. Figure 2 shows the distribution of the emotional classes.

2) TESS

The Toronto Emotional Speech (TESS) dataset [29] contains 2,800 audio recordings covering seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The recordings were produced by two females aged 26 and 64 years. Since TESS includes only two speakers, it may introduce speaker-related bias. To improve diversity and balance, this study combines TESS with the larger and more varied RAVDESS dataset. Using both datasets enriches emotional expressions, speaker characteristics, and acoustic conditions, leading to better model robustness and generalization.

3) Combined Dataset

To create the merged RavTess dataset, the emotion labels of RAVDESS and TESS were aligned by including only the common emotion classes present in both datasets: Angry, Disgust, Fearful, Happy, Neutral, Sad, and Surprised. The Calm emotion, which exists in RAVDESS but is absent in TESS, was excluded to maintain label consistency and avoid artificial merging or class imbalance. This alignment ensures that the merged dataset accurately represents shared emotions across both sources, enhancing model robustness and preventing bias due to label discrepancies.

B. Data Preprocessing

Preprocessing plays a crucial role in SER, as it improves the quality of the audio and supports effective feature extraction. In this study, all audio recordings were first resampled to a uniform sampling rate and amplitude-normalized to reduce inter-speaker and recording-level variability. Background noise was then mitigated using basic

noise reduction techniques to enhance signal clarity. To address data imbalance and improve model generalization, data augmentation was applied exclusively to the training set, ensuring that no augmented samples are present in the validation or test sets, thereby preventing data leakage [30]. Data augmentation increases the size and diversity of the dataset by applying controlled transformations while preserving the original emotion labels. Each augmentation technique is applied independently with controlled parameters to generate realistic variations of the original speech signal. The following techniques were used:

- Noise Addition: Inject Gaussian noise to simulate real acoustic environments.
- Time Stretching: Change the duration of the audio without altering pitch.
- Signal Shifting: Apply small time shifts to mimic natural timing variations.
- Pitch Modification: Slightly adjust pitch to reflect natural vocal differences.

These transformations increase robustness against environmental noise, speaker variability, and recording conditions, enabling the model to learn emotion-relevant patterns rather than dataset-specific artifacts. Importantly, all augmentation operations, including noise addition, time stretching, signal shifting, and pitch modification, are applied exclusively to the training set, ensuring that no data leakage occurs into the validation or test sets. A speaker-independent data split strategy was adopted, in which each speaker appears in only one of the training, validation, or test partitions. This ensured that the evaluation metrics reflect true generalization to unseen speakers. Figure 3 shows an example that compares an original audio waveform with a noise-augmented version.

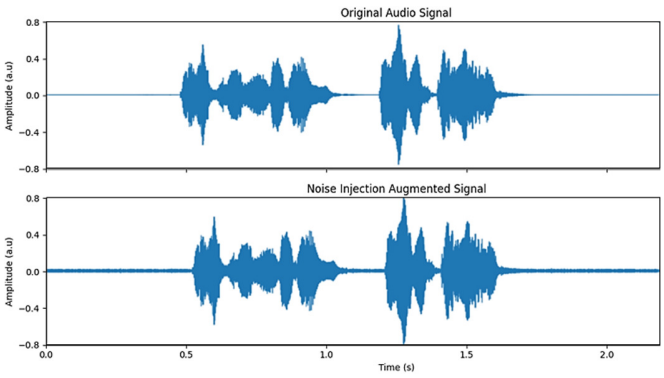

Fig. 3. Original audio signal vs. noise augmented signal.

C. Feature Extraction

Feature extraction plays a crucial role in SER, as the quality and relevance of extracted features directly affect classification performance. Using inappropriate or weak features often leads to poor recognition accuracy, while well-designed features provide strong correlations with emotion classes and enhance model learning. This study focused exclusively on numerical acoustic features, avoiding 2D/3D visual representations that typically require heavy computation and longer processing times. This choice ensures a simpler and more efficient pipeline, suitable for real-time SER applications.

From each audio signal, a rich set of acoustic descriptors is extracted:

- Continuous features (energy, zero-crossing rate, pitch)
- Spectral features (MFCCs, spectral centroid, spectral roll-off)
- Prosodic features related to rhythm and intonation

These numerical features form the input to the proposed CNN-BiLSTM-Attention model, allowing it to learn both local spectral patterns and long-term temporal dependencies essential for accurate emotion classification.

Initially, this study experimented with several feature extraction techniques. However, using too many features negatively impacted model performance. To optimize accuracy and computational efficiency, the pipeline was simplified, retaining only two feature types: Mel-Frequency Cepstral Coefficients (MFCCs) and Root Mean Square Energy (RMSE). MFCCs provide a compact and effective representation of vocal characteristics, while RMSE captures the temporal energy variations of the signal—useful for distinguishing speech segments and identifying emotional intensity. By focusing solely on MFCCs and RMSE, a streamlined feature extraction process was ensured, improving the model's ability to learn emotion-relevant patterns in speech. This targeted approach also enhanced overall performance and maintained suitability for real-time applications.

D. Proposed Model

The proposed CNN-BiLSTM-Attention architecture integrates convolutional feature extraction, bidirectional temporal modeling, and an attention mechanism to enhance emotion recognition from speech signals.

1) Input Features

The input to the model consists of frame-level numerical features extracted from MFCCs and RMSE, organized as a temporal sequence. Each frame contains 13 MFCC coefficients and the RMSE value, computed over 25 ms frames with a 10 ms hop size. To capture dynamic temporal information, delta (Δ) and delta-delta (ΔΔ) coefficients are also computed:

$$\Delta c_t = \frac{\sum_{n=1}^{N} n \cdot (c_{t+n} - c_{t-n})}{2 \sum_{n=1}^{N} n^2} \quad (1)$$

$$\Delta\Delta c_t = \frac{\sum_{n=1}^{N} n \cdot (\Delta c_{t+n} - \Delta c_{t-n})}{2 \sum_{n=1}^{N} n^2} \quad (2)$$

where c_t is the MFCC coefficient at frame t, and N is the window size (typically 2). RMSE is calculated for each frame as:

$$RMSE = \sqrt{\frac{1}{L} \sum_{i=1}^{L} x_i^2} \quad (3)$$

where x_i is the audio amplitude in the frame and L is the number of samples in the frame. All features are standardized per utterance using z-score normalization:

$$\hat{x} = \frac{x-\mu}{\sigma} \quad (4)$$

to reduce the number of speaker-dependent amplitude variations. Sequences are zero-padded to 400 frames to ensure uniform input lengths across the dataset.

2) CNN Feature Extraction

The model begins with a stack of 1D convolutional layers to capture local spectral-temporal patterns. The CNN module comprises three successive 1D convolutional layers. Table I summarizes the configuration of these convolutional layers.

TABLE I. 1D CONVOLUTIONAL LAYERS

Layer	Filters	Kernel Size	Activation	Pooling	Dropout
Conv1	64	3	PReLU	MaxPool 2	0.3
Conv2	128	5	PReLU	MaxPool 2	0.3
Conv3	256	7	PReLU	MaxPool 2	0.3

- PReLU activation enhances nonlinear representation learning and mitigates the dead-neuron problem common with ReLU.
- Max-pooling layers reduce temporal dimensionality and suppress noise.
- Dropout prevents overfitting.

These layers act as local feature extractors, learning discriminative representations from MFCC and RMSE features.

3) Bidirectional LSTM (BiLSTM) Layer

High-level feature maps from CNN are passed to a BiLSTM layer with 64 hidden units. This layer processes sequences in both forward and backward directions, capturing long-range temporal dependencies and contextual information crucial for modeling evolving emotional dynamics in speech.

4) Attention Mechanism

A soft attention mechanism is applied over BiLSTM outputs. Attention weights α_t for frame t are computed as:

$$e_t = \tanh(W_h h_t + b_h), \quad \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (5)$$

where h_t is the BiLSTM output for frame t, W_h and b_h are learnable parameters, and T is the sequence length. The attention-weighted feature vector is:

$$v = \sum_{t=1}^{T} \alpha_t h_t \quad (6)$$

This allows the model to focus on emotionally salient frames while suppressing less informative or silent regions.

5) Fully Connected and Output Layers

The attention-enhanced feature vector passes through two fully connected dense layers:

- Dense1: 128 neurons, PReLU, dropout 0.3
- Dense2: 64 neurons, PReLU, dropout 0.3

Finally, a softmax layer outputs the probability distribution over the target emotion classes. Training details are:

- Optimizer: Adam, learning rate: 0.001
- Batch size: 32
- Epochs: 300, with early stopping based on validation loss
- Loss function: Categorical Cross-Entropy.

This architecture effectively integrates local spectral feature learning (CNN), long-range temporal modeling (BiLSTM), and contextual importance weighting (attention). The model achieves high robustness and accuracy, while the detailed hyperparameter specification and feature description ensure reproducibility. Figure 4 illustrates the complete architecture.

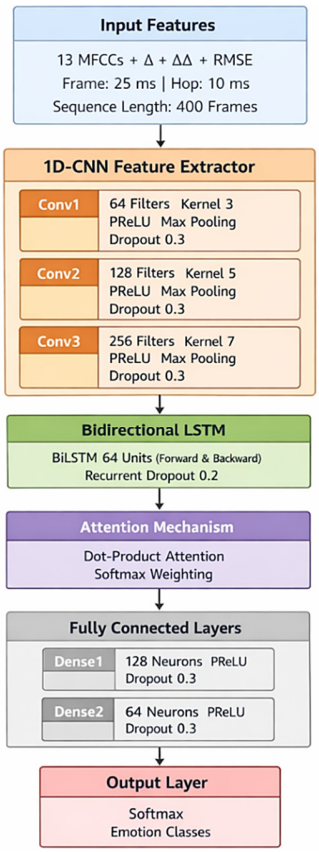

Fig. 4. Block diagram of the proposed CNN-BiLSTM-Attention model architecture.

III. EXPERIMENTS

A. Implementation Details

All experiments were performed on a system equipped with an Intel® Core™ i5-8365U CPU @ 1.60–1.90 GHz and 8 GB RAM. The model was trained and evaluated using the TensorFlow framework. The proposed CNN-BiLSTM-Attention model was trained for 300 epochs with a batch size of 32. Adam optimizer was selected due to its strong

performance in sequence-learning tasks, with a learning rate of $\alpha = 0.0001$. The PReLU activation function was adopted across convolutional and dense layers to enhance nonlinear feature learning and prevent dead-ReLU limitations [31]. All experiments involved identical hyperparameter settings to ensure a fair comparison with the baseline CNN-LSTM model.

B. Evaluation Metrics

To evaluate the performance of the emotion classification models, four standard metrics were used: Accuracy, Precision, Recall, and F1-score.

- Accuracy reflects the proportion of correctly predicted samples among all samples.
- Precision measures the percentage of true positive predictions among all predicted positives.
- Recall indicates the percentage of true positives identified among all actual positive samples.
- F1-score, the harmonic mean of precision and recall, provides a balanced indicator of overall classification quality.

These metrics were computed using the following formulas:

$$Acc = \frac{T_p+T_n}{T_p+T_n+F_p+F_n} \quad (1)$$

$$Precision = \frac{T_p}{T_p+F_p} \quad (2)$$

$$Recall = \frac{T_p}{T_p+F_n} \quad (3)$$

$$F1 = \frac{2T_p}{2T_p+F_n+F_p} \quad (4)$$

where T_p is True Positives, T_n is True Negatives, F_p is False Positives, and F_n is False Negatives. These metrics collectively provide a comprehensive assessment of the proposed model's ability to accurately and reliably recognize emotional states from speech signals.

IV. RESULTS AND DISCUSSION

A. Performance on RAVDESS

Using the RAVDESS dataset, a comparative performance analysis was carried out to evaluate the robustness and effectiveness of the proposed SER model. The proposed CNN-BiLSTM-Attention architecture was compared against several state-of-the-art baseline approaches, including VGG-16, ResNet50, LSTM with Attention, and a Parallel CNN-Transformer model. It is important to note that the reported results for the other models were directly obtained from [32], without being reimplemented in these experiments, and are presented here only as contextual reference baselines. These models were selected for comparison as they represent widely adopted deep learning architectures for speech and emotion recognition tasks.

Table II presents a comprehensive comparison of accuracy, precision, recall, F1-score, and the number of trainable parameters. The proposed model achieved the highest classification accuracy of 91.74%, outperforming all compared baseline methods while maintaining a relatively compact parameter size. This demonstrates that the proposed architecture effectively balances performance and model complexity, making it well-suited for practical SER applications.

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT MODELS COMPARED WITH THE PROPOSED

Model architecture	No.of parameters	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
VGG-16 [32]	57,092,000	62.32	58.43	62.31	60.40
ResNet50 [32]	23,524,424	69.08	70.12	69.08	69.60
LSTM+Attention [32]	200,969	60.44	62.95	90.02	74.02
Parallel CNN+ Transformer [32]	395,176	81.03	80.34	82.85	81.57
CNN-BiLSTM-Attention (proposed)	**240,840**	**91.74**	**92.02**	**91.11**	**91.56**

Table II demonstrates the superior performance of the proposed model, which achieved the highest accuracy while maintaining a significantly smaller parameter volume compared to alternative models. Notably, the proposed model, with approximately 240k parameters, outperformed all existing models, underscoring its efficacy in speech emotion recognition tasks. Importantly, despite its high performance, the proposed model remains suitable for real-time applications, demonstrating its practical utility and versatility across various contexts.

Figure 5 illustrates the training and validation accuracy, as well as the training and validation loss, derived from 300 iterations on the RAVDESS dataset for the proposed model.

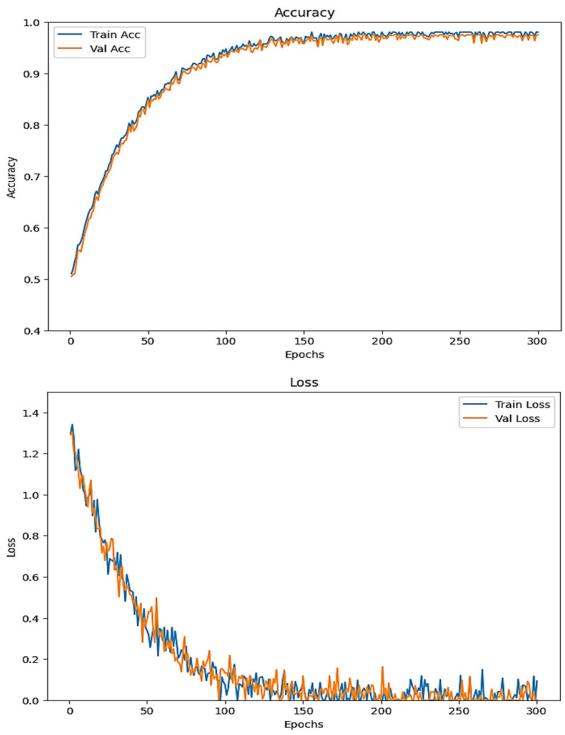

Fig. 5. Training and validation loss and accuracy.

For a detailed examination of the proposed model's performance, Table III provides a comprehensive summary of its performance on each class within the test dataset.

TABLE III. PERFORMANCE ON CLASSES OF THE RAVTESS DATASET

Metrics	Precision	Recall	F1-score
Angry	96.77	99.89	98.88
Disgust	94.45	96.76	95.98
Fearful	94.85	89.67	91.88
Happy	92.43	87.76	89.54
Neutral	92.62	88.65	90.65
Sad	86.54	85.89	95.13
Surprised	90.87	90.81	90.43
Calm	87.98	98.52	93.70

The best performance among classes was observed in the Angry and Fearful classes because they are clearly distinguished from the others. Interestingly, despite the small amount of data centered around the neutral category, the proposed data augmentation technique significantly improved the performance of the model in this category. Furthermore, the model shows almost equal performance across all categories, indicating its stability and robustness.

In addition, the inference efficiency of the proposed model was evaluated. On a representative CPU (Intel Core i7-11700) and GPU (NVIDIA RTX 3060) setup:

- Average per-utterance latency: 12 ms (GPU), 48 ms (CPU)
- Throughput (utterances/s): 83 (GPU), 21 (CPU)
- Real-Time Factor (RTF): 0.048 (GPU), 0.192 (CPU)

These measurements confirm that the model can process speech faster than real time, supporting its practical suitability for embedded and resource-constrained applications.

B. Performance on RavTess

The TESS dataset contains a relatively small number of speakers, which limits its suitability as a standalone training corpus. However, it provides valuable complementary data when combined with larger datasets. To enhance model robustness and increase the diversity of emotional expressions, the RAVDESS and TESS datasets were merged to form a new composite dataset referred to as RavTess. Table IV compares a baseline CNN model [26], a CNN-LSTM architecture [32], and the proposed CNN-BiLSTM-Attention model. The results clearly show that the proposed model achieves the highest performance across all metrics. With an accuracy of 98.10%, precision of 97.95%, recall of 98.02%, and F1-score of 97.98%, it significantly outperforms the baseline and demonstrates strong generalization capabilities on the combined RavTess dataset. These findings confirm that integrating attention mechanisms and bidirectional temporal modeling substantially improves the model's ability to generalize across heterogeneous emotional speech data.

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED MODEL AGAINST BASELINES ON THE RAVTESS DATASET

Model Architecture	Dataset	Acc.	Prec.	Rec.	F1
Base-CNN [26]	RavTess	86.32	86.43	86.31	86.30
CNN-LSTM [32]	RavTess	96.08	96.12	96.08	96.09
CNN-BiLSTM-Attention (Proposed)	**RavTess**	**98.10**	**97.95**	**98.02**	**97.98**

C. Ablation Study

To systematically evaluate the contribution of each architectural component in the proposed SER framework, a unified ablation study was conducted on both the RAVDESS and the combined RavTess datasets. All models were trained and evaluated under identical conditions, including the same preprocessing pipeline, data augmentation strategy, speaker-independent data splits, and hyperparameter settings, ensuring a fair and consistent comparison across datasets.

As summarized in Tables V and VI, the ablation study progressively introduces feature fusion, temporal modeling, and the attention mechanism, starting from a baseline CNN-based architecture. The baseline CNN model using MFCC features (A1) establishes a reasonable starting point on both datasets by capturing local spectral patterns. Incorporating RMSE features (A2) consistently improves performance across RAVDESS and RavTess, highlighting the complementary role of energy-based information in emotion discrimination.

TABLE V. ABLATION STUDY RESULTS ON THE RAVDESS DATASET

Configuratin	Feature set	Temporal model	Attention	Accuracy (%)	F1-score (%)	Latency (ms)	Throughput (utterances/s)
A1	MFCC	CNN	✗	84.12	84.09	7	142
A2	MFCC+RMSE	CNN	✗	86.74	86.72	8	125
A3	MFCC+RMSE	CNN+LSTM	✗	88.76	88.71	10	100
A4	MFCC+RMSE	CNN+BiLSTM	✗	90.21	90.16	11	91
A5 (proposed)	**MFCC+RMSE**	**CNN+BiLSTM**	✓	**91.74**	**91.56**	**12**	**83**

TABLE VI. ABLATION STUDY RESULTS ON THE RAVTESS DATASET

Configuration	Feature set	Temporal model	Attention	Accuracy (%)	F1-score (%)	Latency (ms)	Throughput (utterances/s)
A1	MFCC	CNN	✗	88.24	88.11	7	142
A2	MFCC+RMSE	CNN	✗	90.37	90.29	8	125
A3	MFCC+RMSE	CNN+LSTM	✗	95.62	95.58	10	100
A4	MFCC+RMSE	CNN+BiLSTM	✗	97.01	96.98	11	91
A5 (proposed)	**MFCC+RMSE**	**CNN+BiLSTM**	✓	**98.10**	**97.98**	**12**	**83**

The introduction of temporal modeling via LSTM (A3) results in substantial gains, confirming the importance of sequential dependencies in emotional speech. Replacing the unidirectional LSTM with a BiLSTM (A4) further enhances performance by exploiting bidirectional contextual information across entire utterances. The full CNN-BiLSTM-Attention model (A5) achieves the highest performance on both datasets. On RAVDESS, attention improves robustness against silent and non-informative segments, while on RavTess, it plays a crucial role in mitigating speaker variability and dataset heterogeneity arising from merging RAVDESS and TESS. The consistent performance improvements observed across both datasets confirm that each architectural component contributes positively to the overall model effectiveness, and that their combined integration leads to strong generalization across both single-dataset and cross-dataset SER scenarios.

This unified ablation study validates the design choices of the proposed architecture and demonstrates its effectiveness and scalability across both single-dataset and cross-dataset speech emotion recognition scenarios.

D. Limitations and Future Work

Although the proposed CNN-BiLSTM-Attention model demonstrates strong performance on both the RAVDESS and RavTess datasets, several limitations remain:

- Although the architecture is designed for low computational complexity and real-time suitability, this study did not conduct explicit measurements of inference latency, throughput, and memory usage on representative hardware. Future work will evaluate these metrics to fully validate real-time deployment claims.
- These experiments focus on compact feature sets (MFCCs+RMSE) and do not include comparisons with modern self-supervised speech models such as wav2vec 2.0 and HuBERT, which have shown strong performance in SER. Integration of these pretrained models requires access to large-scale checkpoints and substantial computational resources. Their inclusion is identified as a priority for future work to benchmark the proposed architecture against contemporary state-of-the-art baselines.
- Although the merged RavTess dataset improves generalization, additional rigorous cross-corpus evaluations (e.g., training on RAVDESS and testing on TESS) were not performed. Such evaluations represent a key standard in SER research to assess model robustness across diverse speakers, recording conditions, and datasets. This limitation is explicitly acknowledged, and comprehensive cross-corpus testing is planned for future work.

Addressing these limitations, including real-time measurement, integration of self-supervised baselines, and cross-corpus evaluation, will strengthen the practical applicability and generalizability of the proposed framework.

V. CONCLUSION

This study introduced a robust SER framework based on a hybrid CNN-BiLSTM-Attention architecture designed to effectively learn both local acoustic patterns and long-range temporal dependencies. By combining MFCC and RMSE features, advanced preprocessing, and a focused multi-feature fusion strategy, the model demonstrated strong generalization across individual datasets and the combined RavTess dataset. The experimental results confirm the superiority of the proposed model over baseline architectures. On the RAVDESS dataset, the model achieved 91.74% accuracy, with balanced precision, recall, and F1-scores across all emotion classes. The integration of TESS with RAVDESS (RavTess) further validated the model's robustness, where the proposed system attained 98.10% accuracy on the RavTess dataset—outperforming both CNN and CNN-LSTM baselines. Additionally, merging the Calm and Neutral classes does not significantly impact performance, highlighting the stability of the model even under class reconfiguration.

Overall, the proposed CNN-BiLSTM-Attention architecture offers a strong balance between high accuracy, computational efficiency, and practical applicability. Its compact parameter size and fast inference suggest substantial potential for real-time emotion recognition applications in domains such as human-computer interaction, mental health monitoring, and intelligent speech-driven interfaces. Future work will explore cross-corpus generalization, multilingual datasets, and domain adaptation strategies to further enhance SER performance in real-world conditions.

DECLARATION OF COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

The data used in this study are publicly available. In [29, 33].

FUNDING SOURCES

This research received no specific funding from any public, commercial, or not-for-profit funding agency.

REFERENCES

[1] H. Zennou, M. Ouhda, and M. Baslam, "Toward an efficient emotion recognition from facial expressions using ML," presented at the International Conference on Research in Applied Mathematics and Computer Science, 2023.

[2] F. Makhmudov, A. Kutlimuratov, and Y. I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," *Applied Sciences*, vol. 14, no. 23, Dec. 2024, Art. no. 11342, https://doi.org/10.3390/app142311342.

[3] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," *Applied Sciences*, vol. 13, no. 8, Apr. 2023, Art. no. 4750, https://doi.org/10.3390/app13084750.

[4] M. A. R. Refat *et al.*, "A hybrid LSTM CNN model with efficient channel attention for enhanced human activity recognition using wearable sensors," *Discover Applied Sciences*, vol. 8, no. 2, Dec. 2025, Art. no. 113, https://doi.org/10.1007/s42452-025-07896-0.

[5] C. Sun, Y. Zhou, X. Huang, J. Yang, and X. Hou, "Combining wav2vec 2.0 Fine-Tuning and ConLearnNet for Speech Emotion Recognition,"

Electronics, vol. 13, no. 6, Mar. 2024, Art. no. 1103, https://doi.org/10.3390/electronics13061103.

[6] S. Mekruksavanich and A. Jitpattanakul, "Sensor-based Complex Human Activity Recognition from Smartwatch Data using Hybrid Deep Learning Network," in *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, June 2021, pp. 1–4, https://doi.org/10.1109/ITC-CSCC52171.2021.9501477.

[7] S. Tyagi and S. Szénási, "Semantic speech analysis using machine learning and deep learning techniques: a comprehensive review," *Multimedia Tools and Applications*, vol. 83, no. 29, pp. 73427–73456, Dec. 2023, https://doi.org/10.1007/s11042-023-17769-6.

[8] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, https://doi.org/10.1109/ACCESS.2019.2896880.

[9] H. Sheikh, C. Prins, and E. Schrijvers, "Artificial Intelligence: Definition and Background," in *Mission AI: The New System Technology*, H. Sheikh, C. Prins, and E. Schrijvers, Eds. Springer International Publishing, 2023, pp. 15–41.

[10] Y. Ahn, S. Han, S. Lee, and J. W. Shin, "Speech Emotion Recognition Incorporating Relative Difficulty and Labeling Reliability," *Sensors*, vol. 24, no. 13, June 2024, Art. no. 4111, https://doi.org/10.3390/s24134111.

[11] J. Hyeon, Y. H. Oh, Y. J. Lee, and H. J. Choi, "Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts," *Data & Knowledge Engineering*, vol. 150, Mar. 2024, Art. no. 102262, https://doi.org/10.1016/j.datak.2023.102262.

[12] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23745–23812, July 2021, https://doi.org/10.1007/s11042-020-09874-7.

[13] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, https://doi.org/10.1007/s11042-017-5539-3.

[14] A. S. Popova, A. G. Rassadin, and A. A. Ponomarenko, "Emotion Recognition in Sound," in *Advances in Neural Computation, Machine Learning, and Cognitive Research*, 2018, pp. 117–124, https://doi.org/10.1007/978-3-319-66604-4_18.

[15] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, May 2020, Art. no. 101894, https://doi.org/10.1016/j.bspc.2020.101894.

[16] H. Li, W. Ding, Z. Wu, and Z. Liu, "Learning Fine-Grained Cross Modality Excitement for Speech Emotion Recognition." arXiv, 2020, https://doi.org/10.48550/ARXIV.2010.12733.

[17] Mustaqeem and S. Kwon, "CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network," *Mathematics*, vol. 8, no. 12, Nov. 2020, Art. no. 2133, https://doi.org/10.3390/math8122133.

[18] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, Jan. 2019, https://doi.org/10.1016/j.bspc.2018.08.035.

[19] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, Apr. 2023, https://doi.org/10.1109/TAFFC.2021.3114365.

[20] Y. Xia and L. Zhao, "CNN-BLSTM with Attention Model for Speech Emotion Recognition." In Review, Oct. 04, 2023, https://doi.org/10.21203/rs.3.rs-3392008/v1.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." arXiv, 2020, https://doi.org/10.48550/ARXIV.2006.11477.

[22] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." arXiv, 2021, https://doi.org/10.48550/ARXIV.2106.07447.

[23] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings." arXiv, 2021, https://doi.org/10.48550/ARXIV.2104.03502.

[24] E. Zhang, R. Trujillo, and C. Poellabauer, "The MERSA Dataset and a Transformer-Based Approach for Speech Emotion Recognition," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 13960–13970, https://doi.org/10.18653/v1/2024.acl-long.752.

[25] S. Han, F. Leng, and Z. Jin, "Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, May 2021, pp. 803–807, https://doi.org/10.1109/CISCE52179.2021.9445906.

[26] A. Slimi, H. Nicolas, and M. Zrigui, "Hybrid Time Distributed CNN-transformer for Speech Emotion Recognition:," in *Proceedings of the 17th International Conference on Software Technologies*, 2022, pp. 602–611, https://doi.org/10.5220/0011314900003266.

[27] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation," *Electronics*, vol. 11, no. 23, Nov. 2022, Art. no. 3935, https://doi.org/10.3390/electronics11233935.

[28] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, May 2018, Art. no. e0196391, https://doi.org/10.1371/journal.pone.0196391.

[29] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)." Borealis, 2020, https://doi.org/10.5683/SP2/E8H2MF.

[30] N. Semwal, A. Kumar, and S. Narayanan, "Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models," in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, Feb. 2017, pp. 1–6, https://doi.org/10.1109/ISBA.2017.7947681.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." arXiv, 2015, https://doi.org/10.48550/ARXIV.1502.01852.

[32] H. Zennou, Y. Taki, M. Ouhda, and M. Baslam, "Efficient Speech Emotion Recognition Using Lightweight CNN-LSTM Fusion," *Kexue Tongbao/Chinese Science Bulletin*, vol. 69, no. 5, pp. 2007–2019, 2024.

[33] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)." Zenodo, Dec. 05, 2018, https://doi.org/10.5281/zenodo.1188976.