Asian Journal of Electrical Sciences
ISSN: 2249-6297 (P); 2583-7923 (E)
Vol.14 No.2, 2025, pp.21-27

© Centre for Research and Innovation
www.crijournals.org
DOI: https://doi.org/10.70112/ajes-2025.14.2.4286

Interpretable Speech Emotion Recognition: A Comparative Study of BiLSTM Temporal Attention and Transformer-Based Multi-Head Self-Attention

Rexcharles Enyinna Donatus
Africa Centre of Excellence on Technology Enhanced Learning (ACETEL), National Open University of Nigeria, Nigeria
Department of Aerospace Engineering, Air Force Institute of Technology, Kaduna, Nigeria
E-mail: charlly4eyims@yahoo.com

(Received 22 June 2025; Revised 24 September 2025; Accepted 13 October 2025; Available online 20 October 2025)

Abstract **- Speech Emotion Recognition (SER) is an important area of affective computing that enables machines to understand and respond to human emotions. However, many deep learning approaches that achieve high accuracy provide limited insight into how predictions are made, which reduces their practical reliability in sensitive domains such as education and healthcare. This study presents a comparative analysis of two attention-based models for SER using the RAVDESS dataset: a Bidirectional Long Short-Term Memory (BiLSTM) network with temporal attention and a Transformer model with multi-head self-attention. Acoustic features were extracted using 40 Mel-Frequency Cepstral Coefficients (MFCCs) together with their first- and second-order derivatives, forming a 120-dimensional input feature vector. Both models were trained and evaluated on identical data splits using accuracy, precision, recall, and F1-score. The BiLSTM with temporal attention achieved an accuracy of 70.14% and F1-score of 68.76%, outperforming the Transformer model, which recorded 51.39% and 48.30%, respectively. Attention weight analysis showed that the BiLSTM model concentrated more effectively on emotionally relevant segments of speech, improving interpretability and performance. The findings suggest that incorporating temporal attention provides a better balance between recognition accuracy and model transparency, supporting the development of reliable and explainable SER systems for real-world human–machine interaction.**

Keywords: **Speech Emotion Recognition, Interpretable Deep Learning, BiLSTM, Temporal Attention, Multi-Head Self-Attention, Transformer, MFCC, RAVDESS**

I. INTRODUCTION

Speech Emotion Recognition (SER) is a key area in affective computing, enabling machines to interpret and respond to human emotions in voice-based interactions [1], [2]. Its applications span multiple domains, including human-computer interaction, education, customer service, and mental health monitoring, where accurate emotional understanding enhances system responsiveness and user experience [3], [4], [5]. Despite these benefits, SER remains a technically complex task due to the subtle nature of emotional expression and significant inter-speaker variability across contexts. A core challenge in SER lies in isolating emotion-relevant features from overlapping acoustic factors such as background noise, prosody, and speaker-specific traits [1], [6]. Extracting robust acoustic features is therefore essential for transforming raw audio into informative, discriminative representations. Among these, Mel-Frequency Cepstral Coefficients (MFCCs) remain the most widely adopted due to their efficiency and perceptual relevance [7], [8]. These features, often combined with prosodic and spectral descriptors, serve as foundational inputs for deep learning-based SER models.

Recent studies have explored hybrid neural architectures such as convolutional neural networks (CNNs), BiLSTMs, and attention mechanisms to enhance classification accuracy and robustness [8], [9], [10]. CNNs have proven useful in extracting localized acoustic patterns, while BiLSTM networks are effective at modeling bidirectional temporal dependencies in speech [11], [12]. Attention mechanisms further improve performance by highlighting emotionally salient frames within an utterance [9], [13], [14].

More recently, Transformer-based models employing multi-head self-attention have shown promise in modeling long-range dependencies in speech [1], [15].While these architectures achieve competitive accuracy, their interpretability remains limited due to complex and diffuse attention patterns. In high-stakes domains like healthcare and education, this lack of transparency poses a barrier to adoption, reinforcing the need for interpretable SER systems[16], [17].

In this context, attention-based BiLSTM models have been favored for their balance between performance and interpretability. For example, [18], [19] proposed an Attentive Time-Frequency Neural Network (ATFNN) that integrates both temporal and frequency attention across BiLSTM and Transformer layers. In another work , [20], Proposed a lightweight Speech Emotion Recognition (SER) architecture that integrates attention-based local feature blocks (ALFBs) to capture high-level relevant feature vectors and a global feature block (GFB) technique to capture sequential, long-term dependencies from speech signals By aggregating these local and global contextual feature vectors, the model effectively captures internal correlations reflecting complex human emotional cues The approach was evaluated using four types of spectral features (mel-frequency cepstral coefficients, mel-spectrogram, root mean square value, and zero-crossing rate) and achieved state-of-the-art performance with high mean accuracies (e.g., 99.65% on TESS, 94.88% on RAVDESS, 98.12% on BanglaSER, 97.94% on SUBESCO, and 97.19% on Emo-DB) across five multi-lingual benchmark datasets through a 5-fold cross-validation strategy .

Similarly, Lu *et al.* , [21] introduced Local to Global Feature Aggregation (LGFA), combining a frame-level and segment-level Transformer to preserve both fine-grained and contextual emotional features, achieving high recognition rates on IEMOCAP and CASIA. Another authors W. Chen *et al*., [22] proposed the Deformable Speech Transformer (DST), a novel and adaptive model for speech emotion recognition. DST dynamically adjusts attention window sizes and positions through a deformable attention mechanism guided by a lightweight decision network, enabling the extraction of multi-granularity emotional cues. Evaluated on the IEMOCAP dataset, DST achieved a weighted accuracy of 71.8% and an unweighted accuracy of 73.6%, surpassing existing methods. On the MELD dataset, it attained a weighted F1 score of 48.8%, outperforming current state-of-the-art systems. These results underscore DST's effectiveness and robustness in emotion recognition from speech.

Building on these advances, our study aims to systematically compare two SER architectures: one based on BiLSTM with temporal attention and another based on a Transformer with multi-head self-attention. While both leverage MFCCs as input features, they differ fundamentally in how they model temporal dependencies and in their interpretability. Our goal is to examine not only their predictive performance on the RAVDESS dataset but also their transparency through attention weight analysis, thereby contributing to the design of human-centered, explainable SER systems.Motivated by these observations, this study presents a comparative analysis of two attention-based deep learning architectures for SER: a BiLSTM network with temporal attention and a Transformer model employing multi-head self-attention. Using the RAVDESS dataset, this work evaluates both models in terms of classification performance and interpretability, based on MFCC-derived acoustic features. The primary objective is to determine which attention mechanism provides a better balance between predictive accuracy and model transparency, thereby supporting the development of reliable and explainable SER systems for human-centered applications.

II. METHODOLOGY

A. Dataset and Preprocessing

The study employed the RAVDESS dataset, which contains recordings from 24 professional actors (12 male and 12 female) expressing eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprise. Each emotion was recorded at two intensity levels, resulting in a balanced emotional corpus. All audio samples are in high-quality WAV format, sampled at 48 kHz, providing consistent signal quality for acoustic analysis. Before feature extraction, each utterance was normalized to reduce speaker and amplitude variations and segmented into short overlapping frames suitable for spectral feature computation. The dataset was divided into training and testing subsets using a stratified 80/20 split to preserve emotion and speaker distribution.

B. Feature Extraction

Acoustic features were extracted to represent the emotional characteristics of speech in a structured numerical form suitable for deep learning models. Mel-Frequency Cepstral Coefficients (MFCCs) were chosen because they provide a compact and perceptually relevant description of the speech spectrum. Each audio frame, obtained using a 25 ms analysis window with a 10 ms hop size, was represented by 40 MFCCs. To capture short-term temporal dynamics, the first- and second-order derivatives (delta and delta-delta) were appended, producing a 120-dimensional feature vector per frame. The extracted features were normalized on a per-utterance basis to ensure stable model training and consistent scaling across speakers.

C. Model Architecture

Two deep learning architectures were developed to evaluate the relationship between model interpretability and classification performance: a BiLSTM network with temporal attention and a Transformer encoder with multi-head self-attention. Both models use the same input representation based on 120-dimensional MFCC feature sequences.

1. BiLSTM + Temporal Attention: The first model employs a Bidirectional Long Short-Term Memory (BiLSTM) layer with 128 units in each direction to capture contextual dependencies across time. A dropout rate of 0.5 is applied to reduce overfitting. To enhance interpretability, a temporal attention layer is introduced to assign adaptive weights to different time steps based on their emotional relevance. Given a sequence of BiLSTM hidden states $\boldsymbol{h_t}$, the attention mechanism computes attention weights $\boldsymbol{\alpha_t}$ and a context vector $\boldsymbol{c}$ as:

$$\alpha_t = \frac{\exp(W_t h_t + b_t)}{\sum_i \exp(W_i h_i + b_i)}, \quad c = \sum_t \alpha_t h_t \qquad (1)$$

where W_t and b_t are learnable parameters. The resulting context vector c summarizes the sequence by emphasizing emotionally salient segments. This vector is passed to a dense layer with Softmax activation to classify each utterance into one of eight emotion categories. The total number of trainable parameters is approximately 257,453.

2. Transformer with Multi-Head Self-Attention: The second architecture is based on a Transformer encoder structure designed to capture global dependencies across the entire sequence. The input features are first normalized, followed by a multi-head self-attention block that models temporal relationships in parallel. A position-wise feed-forward network with 128 units and a dropout rate of 0.5 is applied to enhance generalization. The output is aggregated over time using a GlobalAveragePooling1D layer, and a final dense Softmax layer generates the emotion predictions. This model has approximately 140,768 trainable parameters, making it more compact than the BiLSTM-based architecture.

D. Training Procedure

Both models were implemented and trained in a Google Colab environment utilizing GPU acceleration. The models were optimized using the Adam optimizer with a learning rate of 0.001 and trained using the categorical cross-entropy loss function, suitable for multi-class classification. Training was conducted over a maximum of 50 epochs with a batch size of 32. Dropout regularization (rate = 0.5) was applied after the BiLSTM and attention layers to prevent overfitting. A validation split of 20% was used during training, with early stopping (patience = 5) based on validation loss. This configuration ensured efficient convergence and improved generalization.

E. Architectural Summary of BiLSTM and Transformer Models

The layer-wise architectures of both models are presented below. Tables I and II present the layer-wise architecture, output shapes, and parameter counts for the BiLSTM + Temporal Attention model and the Transformer-based Multi-Head Attention model, respectively. These summaries clarify the structural differences and computational complexity of both models.

TABLE I BILSTM + TEMPORAL ATTENTION MODEL ARCHITECTURE

Layer (Type)	Output Shape	Parameters
Input Layer	(None, 165, 120)	0
Bidirectional (BiLSTM)	(None, 165, 256)	254,976
Dropout	(None, 165, 256)	0
Temporal Attention	(None, 256)	421
Dense (Softmax)	(None, 8)	2,056
Total Parameters		257,453

TABLE II TRANSFORMER-BASED MULTI-HEAD ATTENTION MODEL ARCHITECTURE

Layer (Type)	Output Shape	Parameters
Input Layer	(None, 165, 120)	0
Layer Normalization	(None, 165, 120)	240
Multi-Head Self-Attention	(None, 165, 120)	123,768
Dropout	(None, 165, 120)	0
Layer Normalization	(None, 165, 120)	240
Dense (Position-Wise Feed)	(None, 165, 128)	15,488
Dropout	(None, 165, 128)	0
GlobalAveragePooling1D	(None, 128)	0
Dense (Softmax)	(None, 8)	1,032
Total Parameters		140,768

These tables provide a comprehensive summary of the layer configurations, output dimensions, and parameter counts for both models.

III. RESULTS AND DISCUSSION

A. Quantitative Performance Evaluation

Table III summarizes the performance of the BiLSTM + Temporal Attention and Transformer-based models on the RAVDESS dataset. The BiLSTM model achieved higher accuracy (70.14%) and F1-score (68.76%) than the Transformer model (51.39% and 48.30%, respectively). Similar improvements were observed in precision 69.86% vs. 54.59%) and recall (69.61% vs. 49.14%).

These metrics substantiate the advantage of temporal attention in enhancing emotional feature discrimination, particularly in scenarios with subtle prosodic shifts. The Transformer model's lower scores suggest that while capable of capturing global dependencies, it struggles with short or context-sensitive emotional cues.

TABLE III PERFORMANCE COMPARISON BETWEEN TEMPORAL AND MULTI-HEAD SELF-ATTENTION MODELS ON RAVDESS

Model	Accuracy	Precision	Recall	F1-Score
BiLSTM + Temporal Attention	0.7014	0.6986	0.6961	0.6876
Transformer multi-head self-attention	0.5139	0.5459	0.4914	0.4830

These results demonstrate that the temporal attention mechanism enhances the model's ability to discriminate emotional features, particularly in utterances with subtle prosodic variations. Although the Transformer model captures long-range dependencies, its performance suggests difficulty in identifying short-term or context-sensitive emotional cues. (see Figure 4). The comparative analysis highlights that localized temporal modeling provides more robust emotion discrimination in speech data characterized by variable intensity and duration. Beyond classification accuracy, interpretability remains a crucial aspect of speech emotion recognition, especially in applications such as healthcare, education, and affective computing. The BiLSTM model's temporal attention mechanism provides transparent insight into how different time segments contribute to emotion prediction, thereby reinforcing user trust and supporting human-centered deployment. In contrast, the Transformer's multi-head self-attention, while effective in modeling global dependencies, exhibits less discriminative focus and lower performance in capturing short or context-dependent emotional cues. These findings emphasize that temporal attention not only improves predictive accuracy but also enhances the interpretability of SER models, making them more suitable for real-world use.

B. Training Dynamics and Convergence Analysis

Figures 1 and 2 display the training and validation accuracy and loss curves across epochs for both models. The BiLSTM + Temporal Attention model exhibits a smoother and more gradual improvement in accuracy, with training stabilizing around epoch 30. In contrast, the Transformer model converges quickly but begins to plateau early, reaching its stopping point near epoch 15 due to early stopping. The loss curve for the BiLSTM model shows steady minimization with smaller fluctuations, which suggests better generalization. Overall, these trends indicate that while the Transformer learns faster initially, the BiLSTM architecture offers more stable and reliable training dynamics.

C. Attention Weight Visualization and Interpretability

To assess the interpretability of the BiLSTM + Temporal Attention model, we visualized attention weights assigned to different time steps for representative utterances (Fig.3). These plots reveal the model's focus on temporally localized segments such as pitch inflections, stressed syllables, and pauses that correspond to emotionally salient cues in speech. The sharp peaks in attention highlight the model's capacity to isolate the most relevant portions of the sequence that drive its emotion predictions. Fig. 4 presents the attention weight distribution from the Transformer model with multi-head self-attention. Compared to the focused and interpretable patterns of the BiLSTM model, the Transformer exhibits a more diffuse attention spread across the sequence. Although this pattern shows variation, it lacks the sharp discriminative peaks that aid human interpretation. These visual contrasts underscore the trade-off between global context modeling and interpretability. While the Transformer is capable of capturing broader temporal dependencies, the BiLSTM model provides more transparent decision-making a crucial advantage in sensitive domains such as healthcare, education, and assistive technology. Together, Fig. 3 and 4 offer qualitative validation of the attention mechanisms' behavior and their impact on model explainability in SER tasks.

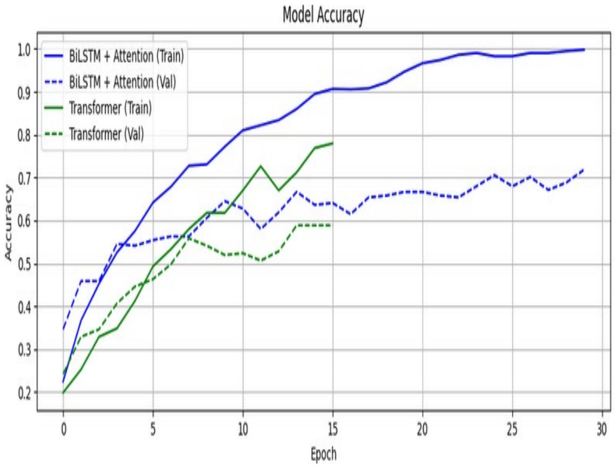

Fig.1 Accuracy vs. Epochs for BiLSTM and Transformer Models

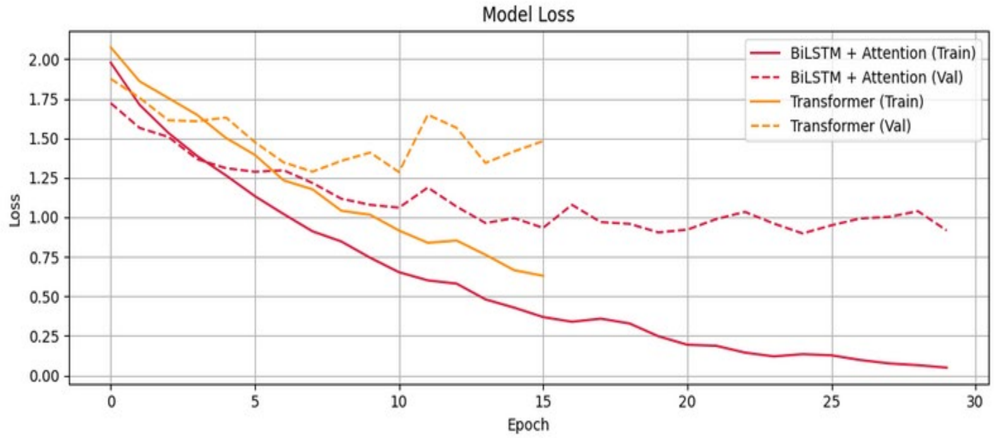

Fig.2 Loss vs. Epochs for BiLSTM and Transformer Models

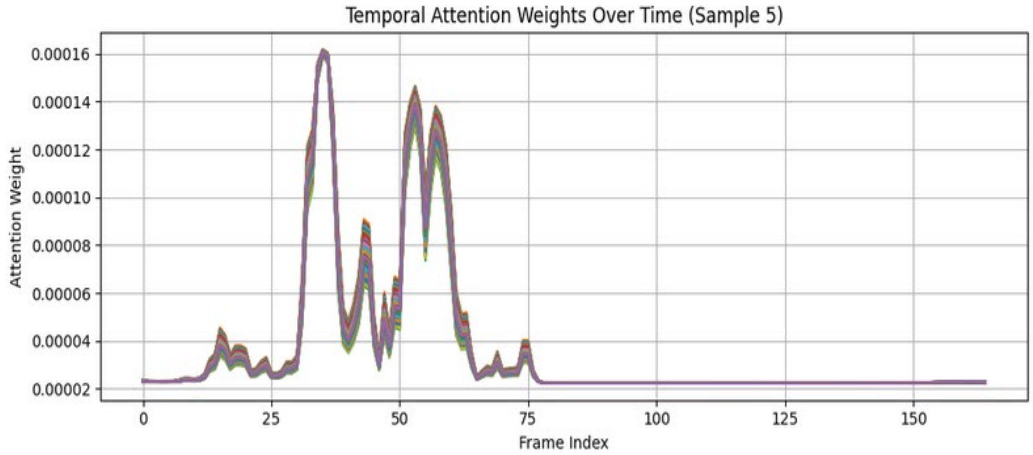

Fig.3 Attention Weight Distribution Over Time in BiLSTM + Temporal Attention Model

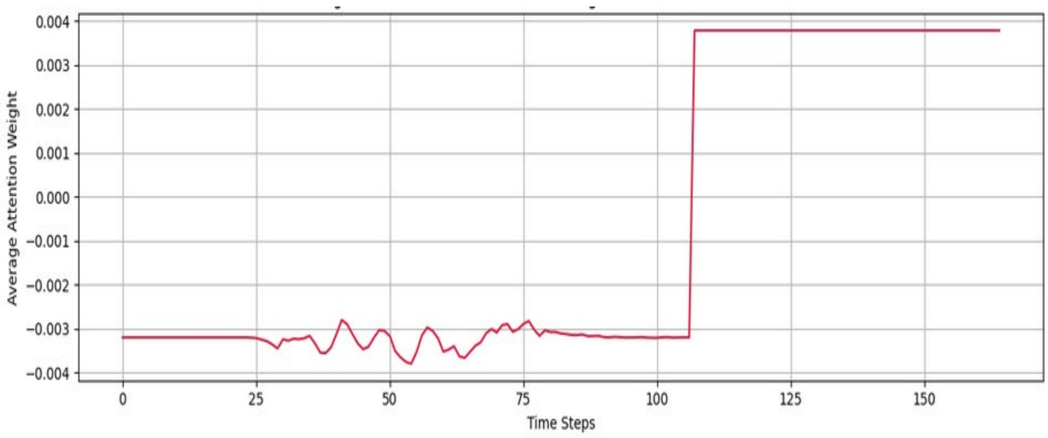

Fig.4 Temporal Distribution of Multi-Head Attention Weights in the Transformer Model

D. Confusion Matrix and Interpretability Analysis

Figures 5 and 6 present the confusion matrices for the BiLSTM and Transformer models, respectively. The BiLSTM model achieves more consistent classification across emotion categories, with fewer confusions among acoustically similar emotions such as calm–sad *and* angry–fearful. In contrast, the Transformer model shows higher misclassification rates within these classes, suggesting difficulty in distinguishing subtle emotional variations. This discrepancy aligns with the attention visualizations, where the BiLSTM demonstrates sharper temporal focus on emotionally salient regions such as pitch fluctuations and intensity peaks thereby enhancing both accuracy and interpretability. These observations confirm that temporal attention provides a dual advantage, delivering superior quantitative performance and clearer qualitative insight in speech emotion recognition tasks.

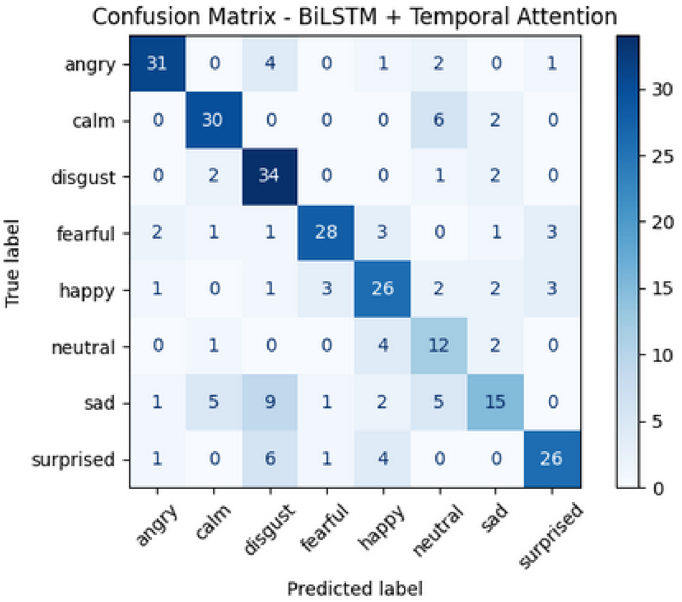

Fig.5 Confusion Matrix – BiLSTM + Temporal Attention

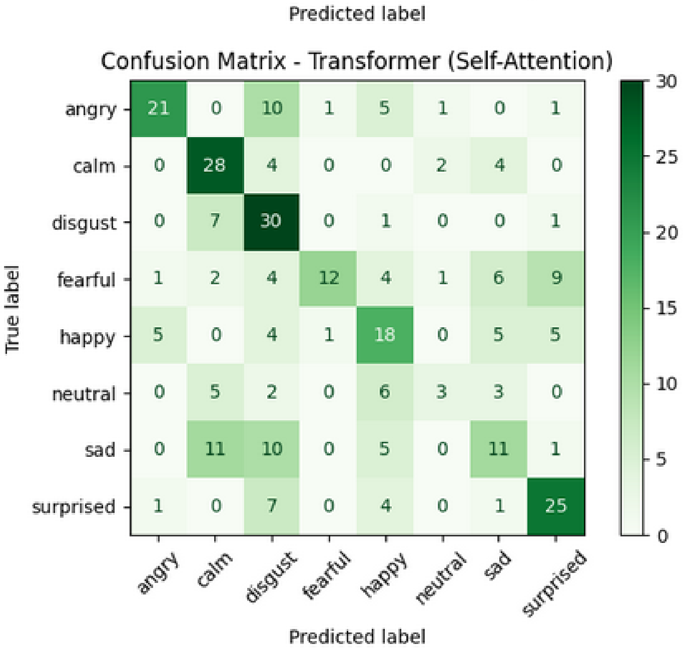

Fig.6 Confusion Matrix – Transformer (Multi-Head Self-Attention)

IV. CONCLUSION

This study presented a comparative evaluation of two attention-based deep learning models for speech emotion recognition (SER) using MFCC-derived acoustic features from the RAVDESS dataset. A Bidirectional LSTM network with temporal attention was compared with a Transformer model employing multi-head self-attention. The BiLSTM + Temporal Attention model consistently outperformed the Transformer across all metrics, achieving an accuracy of 70.14% and an F1-score of 68.76%. Beyond predictive accuracy, the temporal attention mechanism provided transparent interpretability by highlighting emotionally salient regions of speech, a crucial property for deployment in domains such as healthcare, education, and affective computing. In contrast, the Transformer's complex attention patterns captured broader temporal context but were less effective for short or context-dependent utterances and offered limited transparency. These findings emphasize that interpretability should be treated as a core design objective complementing accuracy when developing human-centered AI systems for emotion recognition. Future work will extend this framework to multimodal emotion analysis by incorporating facial and physiological cues, and explore hybrid attention strategies that combine the interpretability of temporal attention with the contextual capacity of self-attention. Domain adaptation approaches will also be investigated to enhance robustness under diverse real-world acoustic conditions.

Declaration of Conflicting Interests

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Use of Artificial Intelligence (AI)-Assisted Technology for Manuscript Preparation

The author confirms that no AI-assisted technologies were used in the preparation or writing of the manuscript, and no images were altered using AI.

ORCID

Rexcharles Enyinna Donatus http://orcid.org/0009-0007-1557-4161

REFERENCES

[1] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "Recognition: Speech, text, and face," *MDPI*, Basel, Switzerland, pp. 1–33, 2023.

[2] R. E. Donatus, B. L. Pal, I. H. Donatus, and U. O. Chiedu, "Comparative analysis of spectrogram and MFCC representations for speech emotion recognition using machine learning," vol. 13, no. 2, pp. 41–47, 2024.

[3] Z. Yang, Z. Li, S. Zhou, L. Zhang, and S. Serikawa, "Speech emotion recognition based on multi-feature speed rate and LSTM," *Neurocomputing*, vol. 601, p. 128177, 2024, doi: 10.1016/j.neucom.2024.128177.

[4] S. S. Chandurkar, S. V. Pede, and S. A. Chandurkar, "System for prediction of human emotions and depression level with recommendation of suitable therapy," *Asian J. Comput. Sci. Technol.*, vol. 6, no. 2, pp. 5–12, 2017, doi: 10.51983/ajcst-2017.6.2.1787.

[5] R. E. Donatus, I. H. Donatus, and U. O. Chiedu, "Exploring the impact of convolutional neural networks on facial emotion detection and recognition," *Asian J. Electr. Sci.*, vol. 13, no. 1, pp. 35–45, 2024.

[6] G. Kaur and S. Baghla, "Speech recognition using cross-correlation algorithm intended for noise reduction," *Asian J. Comput. Sci. Technol.*, vol. 7, no. 3, pp. 48–52, 2018, doi: 10.51983/ajcst-2018.7.3.1899.

[7] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech-based human emotion recognition using MFCC," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw.*, 2017, pp. 2257–2260.

[8] N. Kumar, S. Kobir, and R. Ahmed, "Enhanced speech emotion recognition with efficient channel attention guided deep CNN-BiLSTM framework," *arXiv* preprint arXiv:xxxx.xxxxx, 2024.

[9] Y. Xia, "CNN-BLSTM with attention model for speech emotion recognition," pp. 1–14, 2023.

[10] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, 2021, doi: 10.3390/s21227530.

[11] Benzirar, M. Hamidi, and M. F. Bouami, "Conception of speech emotion recognition methods: A review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 37, no. 3, pp. 1856–1864, 2025, doi: 10.11591/ijeecs.v37.i3.pp1856-1864.

[12] H. S. Kumbhar and S. U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network," in *Proc. 5th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, 2019, pp. 1–3, doi: 10.1109/ICCUBEA47591.2019.9129067.

[13] E. Ghaleb, J. Niehues, and S. Asteriadis, "Joint modelling of audio-visual cues using attention mechanisms for emotion recognition," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11239–11264, 2023, doi: 10.1007/s11042-022-13557-w.

[14] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with lightweight deep neural ensemble model using handcrafted features," *Sci. Rep.*, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-95734-z.

[15] P. Karmakar, S. W. Teng, and G. Lu, "Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition," *Intell. Syst. Appl.*, vol. 23, p. 200406, 2024, doi: 10.1016/j.iswa.2024.200406.

[16] P. Kumar, S. Malik, and B. Raman, "Interpretable multimodal emotion recognition using hybrid fusion of speech and image data," *Multimedia Tools Appl.*, vol. 83, no. 10, pp. 28373–28394, 2024, doi: 10.1007/s11042-023-16443-1.

[17] S. Das, N. N. Lønfeldt, A. K. Pagsberg, and L. H. Clemmensen, "Towards interpretable and transferable speech emotion recognition: Latent representation-based analysis of features, methods and corpora," *arXiv* preprint arXiv:2105.02055, 2021. [Online]. Available: http://arxiv.org/abs/2105.02055

[18] C. Lu, W. Zheng, H. Lian, Y. Zong, and C. Tang, "Speech emotion recognition via an attentive time-frequency neural network," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 6, pp. 3159–3168, 2022.

[19] H. Zhang, H. Huang, and H. Han, "Attention-based convolution skip bidirectional long short-term memory network for speech emotion recognition," *IEEE Access*, vol. 9, pp. 5332–5342, 2021, doi: 10.1109/ACCESS.2020.3047395.

[20] Z. Zhang and K. Wang, "Multiple attention convolutional-recurrent neural networks for speech emotion recognition," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interact. Workshops (ACIIW)*, 2022, pp. 1–8, doi: 10.1109/ACIIW57231.2022.10086021.

[21] C. Lu *et al*., "Learning local-to-global feature aggregation for speech emotion recognition," *arXiv* preprint arXiv:2306.01491, 2023.

[22] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "DST: Deformable speech transformer for emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096966.