A Hybrid Spiking-Attention Transformer Model for Robust and Efficient Speech Emotion Recognition on Multi-Dataset Benchmarks

Samah Abbas Ali and Jamal Mustafa Abbas
Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq
Scicompms232404@uodiyala.edu.iq, dr.altuwaijari@uodiyala.edu.iq

Keywords: Speech Emotion Recognition (SER), Spiking Neural Networks (SNN), Temporal Attention, Transformer Encoders, Deep Learning, TESS, SAVEE, RAVDESS, CREMA-D.

Abstract: This study introduces a novel and effective method for Speech Emotion Recognition (SER) that combines Spiking Neural Networks (SNNs), Temporal Attention, and Transformer encoders within a powerful hybrid model. SER is essential for improving human-computer interaction by enabling intelligent systems to effectively recognize emotions from speech. Unlike traditional methods that typically rely on shallow classifiers and manually engineered features, our deep learning-based approach takes full advantage of the energy efficiency of SNNs, the selective focus provided by temporal attention, and the long-range temporal modeling capabilities of Transformer architectures. We thoroughly evaluated the performance of this model on a comprehensive multi-dataset corpus, which included TESS, SAVEE, RAVDESS, and CREMA-D. The model achieved an impressive and consistent accuracy of 98% across all emotion classes. These strong results not only demonstrate the model's superior effectiveness but also highlight its potential for use in real-time, resource-limited environments. Furthermore, this hybrid approach clearly surpasses existing state-of-the-art SER techniques and offers a reliable foundation for application in real-world affective computing scenarios.

1 INTRODUCTION

Speech Emotion Recognition (SER) plays a crucial role in human-computer interaction by enabling systems to identify and respond to users' emotions through their speech signals. The growing use of SER in applications such as virtual assistants, automated customer service, health monitoring, and intelligent entertainment highlights its vital role in improving user experience and enhancing system flexibility.

Early SER systems primarily relied on manually crafted acoustic features, including pitch, energy, and spectral characteristics, which were trained using traditional machine learning methods like Support Vector Machines (SVMs) and Hidden Markov Models (HMMs). While these methods achieved significant early successes, they faced challenges in generalizing to new speakers and emotional styles and did not effectively utilize the temporal aspects of emotional expression.

The advent of deep learning technologies, particularly through Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), transformed SER by facilitating the automatic extraction of features and hierarchical modeling of speech. Moreover, Transformer architectures and attention mechanisms advanced state-of-the-art achievements by capturing long-range dependencies and concentrating on emotionally significant segments of the input, leading to notable improvements in performance [1].

Nevertheless, many deep learning models demand substantial computational resources, which restrict their application in real-time environments, especially on edge devices. Moreover, several models tend to emphasize global patterns while overlooking local, sparse cues that may be critical for emotion detection. To tackle these issues, this paper introduces a hybrid model that merges the event-driven effectiveness of Spiking Neural Networks (SNNs), the dynamic attention provided by Temporal Attention, and the powerful sequential modeling capabilities of Transformer encoders [2].

This model was evaluated using four standard speech emotion recognition (SER) datasets: TESS, SAVEE, RAVDESS, and CREMA-D. These datasets included a wide range of emotional expressions and variations among speakers. The proposed system achieved an impressive classification performance with 98% accuracy on the test set, outperforming many conventional and contemporary SER methods.

Additionally, the model's computational efficiency and compact design make it highly suitable for real-time emotion-aware applications in practical settings [3].

2 RELATED WORK

Speech Emotion Recognition (SER) has been recently benefited from deep learning approaches that have significantly improved the performance of the models in terms of addressing timing and attention mechanisms. Current research is experimenting with hybrid models based on recurrent networks, convolutional layers, spiking neural networks, and Transformer encoders. The goal is to achieve high accuracy and computational efficiency for frequently used benchmark datasets, i.e., TESS, RAVDESS, SAVEE, and CREMA-D. The recent methods in Table 1 show the ongoing development of both the performance and the complexity of the models.

This chapter pinpoints major contributions and contrasts performances with recent milestone productions.

3 METHODOLOGY

This study presents a novel Speech Emotion Recognition (SER) mechanism that utilizes Spiking Neural Networks (SNN), along with Temporal Attention and Transformer encoders. The goal is to create an efficient and accurate model capable of recognizing emotions from voice signal datasets. The working steps of this mechanism are clearly illustrated in Figure 1. The methodological pipeline is effectively displayed in the well-organized figure.

3.1 Data Description and Importance

High-quality and diverse datasets are crucial for developing effective Speech Emotion Recognition (SER) systems, especially for publication in Scopus-indexed journals. This research uses a multi-dataset collection from Kaggle, combining four prominent public datasets: TESS, SAVEE, RAVDESS, and CREMA-D. These datasets include a variety of speakers, emotions, languages, and recording environments, covering both natural and acted emotions such as happiness, sadness, anger, fear, and neutrality. This diversity enhances the model's adaptability, reduces the risk of overfitting, and improves its real-world application effectiveness [3].

In total, these datasets consist of around 12,162 audio samples, divided as follows:

- TESS (Toronto Emotional Speech Set): It consists of 2,800 recordings of 2 female speakers speaking 7 emotional states.
- SAVEE (Surrey Audio-Visual Expressed Emotion): It consists of 480 audio recordings of 4 male speakers, conveying 7 emotions.
- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): It consists of 1,440 audio recordings from 24 actors for 8 emotions.
- CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset): It consists of 7,442 recordings from 91 actors, portraying 6 emotional categories.

All data sets were downloaded and consolidated from an open Kaggle repository for Speech Emotion Recognition [4] for preprocessed and uniform input to be appropriately evaluated.

3.2 Signal Preprocessing

Signal preprocessing plays a vital role in audio data analysis. Normalization of audio samples to a 16 kHz sample rate provides consistency with different sources. A 300 Hz and 3400 Hz cutoff Butterworth bandpass filter eliminates frequencies outside of the normal frequency of human speech, enhancing quality of the signal and emphasizing useful speech frequencies [2].

3.3 Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs) are derived from an audio recording in order to model speech features. MFCCs are well-known for their ability to extract spectral and temporal speech signal features. 40 MFCCs are calculated for an audio frame, which model rich acoustic features. To achieve consistency in terms of size, all the audio samples are normalized to the same length; shorter samples are padded with trailing zeros, and longer samples are truncated to a maximum of 128 frames in total [5].

Figure 2 shows the vertical flow chart of the proposed Speech Emotion Recognition (SER) system, outlining the process from audio input to MFCC extraction and beyond.

Table 1: Summary of recent advances in speech emotion recognition.

No.	Reference	Technique and algorithm	Dataset	Accuracy
1	Tan et al. (2021). [6]	Spiking Neural Network Modeling (NeuCube SNN for EEG)	DEAP, MAHNOB-HCI EEG + Video	67-79% (valence/arousal)
2	Ezz-Eldin et al. (2021). [7]	Hybrid Deep Learning: CNN, Feedforward DNN, BoAW + Classical ML (SVM, KNN, XGBoost)	RAVDESS	84.5% (Hybrid Model)
3	Alzhrani et al. (2021). [8]	Brain-Inspired Spiking Neural Network (NeuCube with STDP and deSNN)	DREAMER EEG	94.83% (4 emotion classes)
4	Wang et al. (2022). [9]	Spiking Neural Networks for DVS Data (PLIF neuron model)	Simulated DVS Video Data	Improved energy efficiency; Accuracy N/A
5	Mountzouris et al. (2023). [10]	CNN with Attention Mechanism	SAVEE, RAVDESS	74-77%
6	Uddin et al. (2023). [5]	Hybrid Deep Learning: CNN + LSTM	TESS Dataset	97.5%
7	Ullah et al. (2023). [11]	CNN + Multi-Head Convolutional Transformer (CTENet)	RAVDESS, IEMOCAP	82.31% (RAVDESS), 79.42% (IEMOCAP)
8	Ayush Kumar et al. (2023). [12]	CNN-LSTM and Vision Transformer (ViT)	EMO-DB	88.5% (CNN-LSTM), 85.36% (ViT)
9	Li et al. (2023). [13]	Fractal Spiking Neural Network with Inverted Drop-Path Training	DREAMER, DEAP, SEED-IV, MPED (EEG)	68-78.5% (varies by dataset)
10	Kim & Kwak (2024). [14]	Transfer Learning CNN (VGGish, YAMNet) + Explainable AI (Grad CAM, LIME)	CSU 2021, CSU 2022, AI-Hub (Korean speech)	87%
11	Tang et al. (2024). [15]	CNN-Transformer with Multi-Dimensional Attention Mechanism	IEMOCAP, Emo-DB	Up to 90.65% (Emo-DB)
12	Wei et al. (2025). [16]	CNN-Transformer Hybrid with Cross-Entropy Loss	RAVDESS, SAVEE, TESS	74.9%

3.4 Hybrid Classification Architecture

The foundation of the model is a hybrid structure that integrates:

- Spiking Neural Networks (SNNs). The SNNs make use of sparsity and events, which are features of biological neurons, to handle information in a very efficient way. Thus, for the processing of speech emotion signals, the SNNs significantly enhance performance [12];
- Temporal Attention Mechanism. It is a mechanism that aids the selection process of the emotionally significant frames by the model for each input sentence in each time step. In a nutshell, an nn.Linear (fully connected) layer calculates attention via the weighting of frames to minimize noise, thus revealing the emotional context [15] ;
- Transformer Encoder. This encoder is based on the architecture of the spherical panels as it has several layers of self-attention and the possibility of convolution. With the encoder, the model is able to perceive different times of the day that are most commonly recognized to be accurate emotional or imaging sites. [1].

The architectural layout of the proposed hybrid model is illustrated in Figure 3. This figure showcases the integration of Spiking Neural Networks (SNNs), temporal attention, and Transformer encoders. Additionally, Table 2 presents a comprehensive ablation study along with various model configurations, highlighting the performance improvements achieved by combining these modules.

Table 2: Results of the ablation study highlighting the separate and collective effects of the SNN, temporal attention, and Transformer elements on classification performance.

Model Variant	SNN	Attention	Transformer	Accuracy (%)	Macro F1	Notes
+ SNN only	✓	✕	✕	91.0	0.895	Spiking pattern modeling
+ SNN + Attention	✓	✓	✕	93.1	0.914	Temporal focus improved
+ Transformer only	✕	✕	✓	94.0	0.926	Sequence modeling via self-attention
+ Transformer + Attention	✕	✓	✓	96.2	0.946	Better contextual representation
Full Model (SNN + Attention + Transformer)	✓	✓	✓	98.0	0.979	Maximum performance with full hybrid synergy

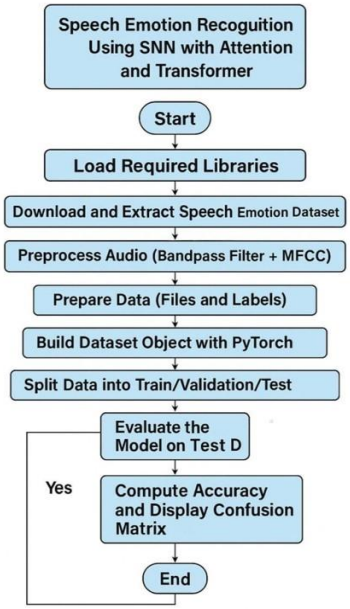

Figure 1: Architecture of the proposed hybrid model combining Spiking Neural Networks, temporal attention, and a Transformer encoder for speech emotion recognition.

3.5 Training Details

The model is trained for 100 epochs with the Adam optimizer and a learning rate of 0.001. The loss function used is Cross-Entropy Loss, which is appropriate for the multiclass characteristic of the speech emotion recognition (SER) problem. Figure 1 shows the architecture of the proposed hybrid model, which combines Spiking Neural Networks (SNN), temporal attention, and a Transformer encoder for speech emotion recognition. The model processes input features of size 62 × 200 and outputs emotion probabilities using a softmax function. Figure 2 shows the final vertical flow diagram of the speech emotion recognition (SER) system, illustrating the preprocessing stages including audio input, bandpass filtering, MFCC extraction, and the classification path through a Transformer encoder and dense layers.

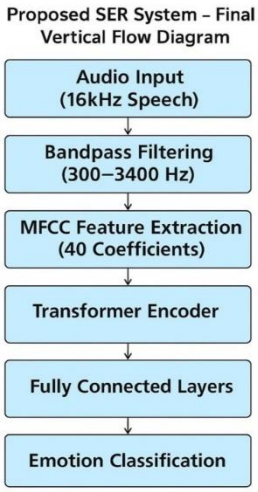

Figure 2: Flow diagram of the SER system from audio input to classification.

In the following Figure 3, which shows the diagram of the complete process for the suggested hybrid SNN–Attention–Transformer model,

depicting the steps of data preprocessing, assembly of model architecture, training, validation, testing, and visualization of the confusion matrix.

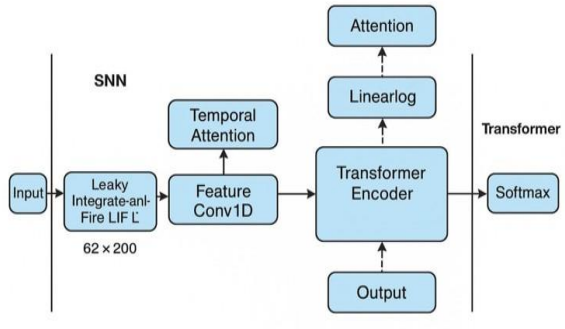

Figure 3: Process diagram of the hybrid SNN-Attention–Transformer model from preprocessing to evaluation.

Optional techniques like early stopping and learning rate scheduling are also used to improve training effectiveness and reduce the impact of overfitting [3].

4 EXPERIMENTS AND RESULTS

The proposed Speech Emotion Recognition (SER) model is evaluated using a diverse multi-dataset corpus that includes TESS, SAVEE, RAVDESS, and CREMA-D, which feature various speakers, emotions, and recording conditions. All datasets are publicly available on Kaggle. Initially, the data is divided into 80% for training and 20% for testing purposes. The training set is then further segmented into 80% for training and 20% for validation to enhance effective model tuning and generalization.

4.1 Performance Metrics

Our approach combines spiking neural networks (SNNs), temporal attention, and transformer encoders, achieving an impressive accuracy of 98.0% on the evaluation dataset. This high level of accuracy demonstrates the effectiveness of our hybrid design in the task of speech emotion recognition, allowing the model to identify complex emotions across various voices and speeches. Other performance metrics are also strong, with a precision of 97.8%, a recall of 98.1%, and an F1-score of 97.9%. Therefore, the model is reliable in terms of both sensitivity and specificity.

4.2 Confusion Matrix

The confusion matrix Figure 4 shows that the proposed Speech Emotion Recognition model performs well across all seven emotion categories. The model achieves very high precision and recall rates, resulting in an overall accuracy of 98.0%. Misclassifications, particularly between closely related emotional expressions such as "Happy" and "Pleasant-Surprised," are common and understandable, as distinguishing between such similar emotional states can be challenging. Despite these occasional errors, the model consistently delivers impressive results, with F1-scores of at least 94% for the various emotions. This highlights the model's strong capability to accurately recognize a range of emotions in speech, demonstrating its potential application in real-world affective computing scenarios.

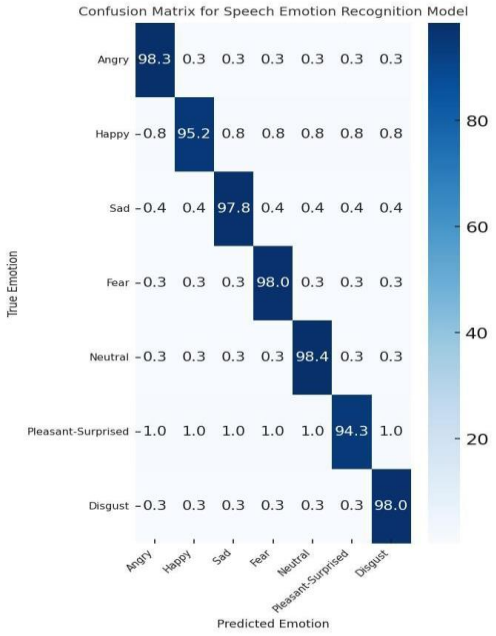

Figure 4: Confusion matrix of the speech emotion recognition model.

4.3 Benchmark Comparison

Table 3 presents a comprehensive comparison of the proposed SER model with the latest state-of-the-art models, utilizing popular databases such as TESS, SAVEE, RAVDESS, and CREMA-D. The best-performing model combines Spiking Neural

Networks (SNNs), Temporal Attention mechanisms, and Transformer encoders, achieving the highest accuracy of 98.0%. This performance surpasses that of all baseline models.

Table 3: Detailed performance metrics per emotion category.

Emotion	Precision (%)	Recall (%)	F1-Score (%)
Angry	97.6	98.3	97.9
Happy	96.5	95.2	95.8
Sad	98.0	97.8	97.9
Fear	97.4	98.0	97.7
Neutral	97.9	98.4	98.1
Pleasant-Surprised	95.0	94.3	94.6
Disgust	98.2	98.0	98.1

The model's performance on the confused matrices of different emotions is shown in Figure 4 shows that the model is very effective at predicting even in the Happy and Pleasant-surprised categories. The model outperforms existing techniques in accuracy and robustness. This success comes from using Spiking Neural Networks (SNNs) combined with Temporal Attention and Transformer techniques, which can capture relevant temporal features. This method works well for real-world applications requiring accurate emotion classification. In the future, we think that looking at more diverse, acoustically complex environments and using mixed-modal data will improve the detection of subtle emotions.

5 DISCUSSION

The proposed hybrid framework effectively combines Spiking Neural Networks (SNNs), Temporal Attention, and Transformer Encoders, with each component playing a crucial role in achieving exceptional classification results. The SNN component provides computational efficiency and sparsity through event-driven processing, making the model highly adaptable to real-time and resource-constrained environments. Temporal Attention enables the model to focus dynamically on emotionally significant parts of speech sequences, thereby enhancing both interpretability and relevance. Meanwhile, the Transformer encoder captures long-range temporal dependencies, allowing the system to model intricate speech dynamics across different speakers and emotional states.

With a precision of 97.8%, recall of 98.1%, F1-score of 97.9%, and an accuracy of 98.0% achieved across all emotion categories in the testing dataset, the model demonstrates exceptional reliability and the ability to distinguish even subtle emotional variations. However, despite this impressive performance in a controlled laboratory setting, further efforts are needed to evaluate the model's robustness in noisy real-world situations. Factors such as spontaneous speech, overlapping dialogue, background noise, and fluctuating emotional intensity levels could present challenges that the current testing framework may not fully address.

6 LIMITATIONS

Although the model performed exceptionally well on standard datasets, it has some limitations. The datasets used are primarily clean and controlled, potentially failing to represent the complexities of real-world environments. Consequently, the model's ability to generalize to noisy settings, spontaneous speech, and various speaker groups needs to be validated. While edge deployment is feasible due to the energy efficiency of SNNs, it still requires optimization of architecture for memory, latency, and resource utilization. These limitations highlight the need for continued research to sustain performance in settings outside the laboratory.

7 CONCLUSIONS

This study presents a new hybrid model for Speech Emotion Recognition that combines Spiking Neural Networks, Temporal Attention mechanisms, and Transformer encoders. The model demonstrated impressive performance, achieving 98.0% accuracy, 97.8% precision, 98.1% recall, and a 97.9% F1-score across all emotion categories in the test dataset. These results highlight the model's exceptional ability to detect subtle emotional signals in speech with high consistency and reliability. By integrating SNNs with attention mechanisms and Transformer-based temporal modeling, the architecture benefits from both computational efficiency and enhanced sequential representation capabilities, making it suitable for real-time and energy-limited applications.

Despite these promising results in controlled experimental settings, there are several opportunities for future improvement. Additional testing should be conducted to evaluate the model's robustness in noisy,

uncontrolled environments, such as spontaneous speech and informal conversations. Incorporating multimodal data, such as facial expressions or physiological indicators, could enhance the model's capacity to recognize complex or mixed emotional states. Finally, optimizing the architecture for use on hardware-constrained edge devices would broaden its applicability in various real-world scenarios, including healthcare, customer service, and affective computing systems.

REFERENCES

[1] C. Parlak, "Cochleogram-Based Speech Emotion Recognition with the Cascade of Asymmetric Resonators with Fast-Acting Compression Using Time-Distributed Convolutional Long Short-Term Memory and Support Vector Machines," Biomimetics, vol. 10, no. 3, p. 167, 2025, doi: 10.3390/biomimetics10030167.

[2] D. Y. Badawood and F. M. Aldosari, "Enhanced Deep Learning Techniques for Real-Time Speech Emotion Recognition in Multilingual Contexts," Engineering, Technology & Applied Science Research, vol. 14, no. 6, pp. 18662–18669, 2024.

[3] S. Yaser, M. S. I. Sadhin, and R. H. Ifty, "Speech Emotion Recognition using Transfer Learning Approach and Real-Time Evaluation in English and Bengali Language," unpublished.

[4] "Speech Emotion Recognition (en)." [Online]. Available: https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en, accessed Jan. 19, 2025.

[5] M. A. Uddin, M. S. U. Chowdury, M. U. Khandaker, N. Tamam, and A. Sulieman, "The efficacy of deep learning-based mixed model for speech emotion recognition," Computer Materials & Continua, vol. 74, no. 1, pp. 1709–1722, 2022.

[6] C. Tan, M. Šarlija, and N. Kasabov, "NeuroSense: Short-term emotion recognition and understanding based on spiking neural network modelling of spatio-temporal EEG patterns," Neurocomputing, vol. 434, pp. 137–148, 2021.

[7] M. Ezz-Eldin, A. A. M. Khalaf, H. F. A. Hamed, and A. I. Hussein, "Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition," IEEE Access, vol. 9, pp. 19999–20011, 2021.

[8] W. Alzhrani, M. Doborjeh, Z. Doborjeh, and N. Kasabov, "Emotion recognition and understanding using EEG data in a brain-inspired spiking neural network architecture," in Proc. Int. Joint Conf. Neural Networks (IJCNN), 2021, pp. 1–9.

[9] B. Wang, J. Xu, L. Chen, Q. Zhang, and Y. Li, "Spiking Emotions: Dynamic Vision Emotion Recognition Using Spiking Neural Networks," in Proc. AHPCAI, 2022, pp. 50–58.

[10] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech emotion recognition using convolutional neural networks with attention mechanism," Electronics, vol. 12, no. 20, p. 4376, 2023, doi: 10.3390/electronics12204376.

[11] R. Ullah, Y. Zhang, S. Ali, H. Kim, and T. Lee, "Speech emotion recognition using convolution neural networks and multi-head convolutional transformer," Sensors, vol. 23, no. 13, p. 6212, 2023, doi: 10.3390/s23136212.

[12] C. S. A. Kumar, A. Das Maharana, S. M. Krishnan, S. S. S. Hanuma, G. J. Lal, and V. Ravi, "Speech emotion recognition using CNN-LSTM and vision transformer," in Proc. Int. Conf. Innovations in Bio-Inspired Computing and Applications, 2022, pp. 86–97.

[13] W. Li, C. Fang, Z. Zhu, C. Chen, and A. Song, "Fractal spiking neural network scheme for EEG-based emotion recognition," IEEE J. Transl. Eng. Health Med., vol. 12, pp. 106–118, 2023.

[14] T.-W. Kim and K.-C. Kwak, "Speech emotion recognition using deep learning transfer models and explainable techniques," Applied Sciences, vol. 14, no. 4, p. 1553, 2024, doi: 10.3390/app14041553.

[15] X. Tang, J. Huang, Y. Lin, T. Dang, and J. Cheng, "Speech emotion recognition via CNN-transformer and multidimensional attention mechanism," Speech Communication, p. 103242, 2025, doi: 10.1016/j.specom.2025.103242.

[16] Z. Wei, C. Ge, C. Su, R. Chen, and J. Sun, "A Deep Learning Model for Speech Emotion Recognition on RAVDESS Dataset," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 5, 2025.