STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, July 2025, pp 340–351.
Published online in International Academic Press (www.IAPress.org)

A Hybrid Approach of Long Short Term Memory and Transformer Models for Speech Emotion Recognition

Tarik AbuAin*

College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia

Abstract Speech emotion recognition (SER) has become a critical component of the next generations of technologies that interact between humans and machines. However, in this paper, we explore the advantage of the hybrid LSTM + Transformer model over the solo LSTM and Transformer models. The proposed method contains the following steps: data loading using benchmark datasets such as the Toronto Emotional Speech Set (TESS), Berlin Emotional Speech Database (EMO-DB), and (SAVEE). Secondly, to create a meaningful representation to preprocess raw audio data, Mel-Frequency Cepstral Coefficients (MFCCs) are used; thirdly, the model's architecture is designed and explained. Finally, we evaluate the precision, recall, F1 score, classification reports, and confusion matrices of the model. The outcome of this experiment based on classification reports and confusion matrices shows that the hybrid LSTM + Transformer model has a remarkable performance on the TESS-DB, surpassing the other models with a 99.64% accuracy rate, while the LSTM model gained 97.50% and the Transformer model achieved 98.21%. For the EMO-DB, the LSTM model achieved the highest accuracy of 73.83%, followed by the hybrid that gained 71.96%, and the Transformer model achieved 70.09%. Lastly, LSTM obtained the highest performance on SAVEE-DB of 65.62% accuracy, followed by the Transformer model which achieved 58.33%, and the hybrid model achieved 56.25%.

Keywords Speech Recognition, Emotion Recognition, Sentiment Analysis, LSTM Model, Transformer Model

DOI: 10.19139/soic-2310-5070-2521

1. Introduction

Making machines understand humans has become the trend and requirement of the network era due to the quick development of artificial intelligence technologies. Emotions are the most critical component of human communication and can be utilized to evaluate paralinguistic human expressions and other things. Therefore, voice signals are an effective means of smoothing the fastest communication between human-computer interaction (HCI) and effectively recognizing human behavior [1]. Speech has emerged as a major area of study for many academics in human–computer interaction because of its practical and useful qualities [2]. Detecting emotional expressions is still considered a challenging part, as the voice state is variable all the time, and thus it is hard to identify it accurately most times [3]. SER is a machine learning (ML) problem in which speech utterances are categorized based on the emotions that underlie them. An overview of the most common classification methods in SER is provided in this chapter. The researchers used a variety of classifiers for SER, but generally no adequate rationale is given to select a specific classification model [4]. This paper is organized as follows: Related works are discussed in the first section, followed by the proposed methodology in the paper along with a hybrid approach model used that combines LSTM and Transform in the second, the experimental findings in the third, the discussion of the experiment in the fourth, and the conclusion in the fifth. The purpose of this study is to unveil the strength of the

*Correspondence to: Tarik AbuAin (Email: t.aboain@seu.edu.sa). College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

hybrid LSTM + Transformer over the solo LSTM and Transformer models across three types of datasets such as TESS, EMO-DB, and SAVEE. Figure 1 below illustrates the workflow of speech recognition systems.

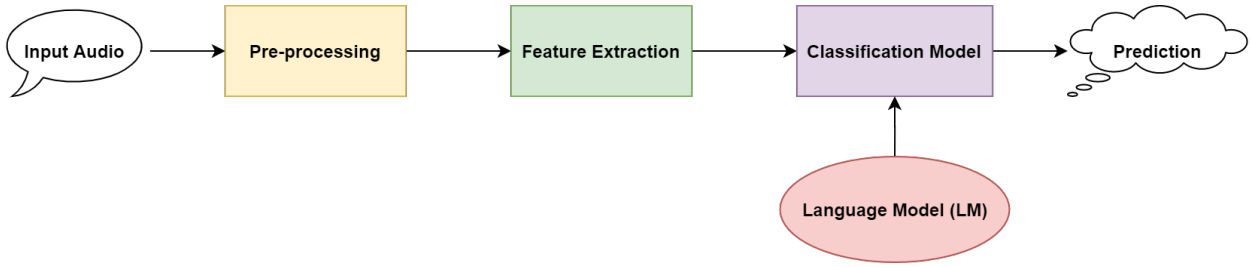

Figure 1. Workflow of speech recognition systems [5]

2. Related Work

There are numerous other recent papers and studies on SER because of its significance in human–computer interaction and the creation of artificial intelligence systems. We have examined the most recent research that is relevant to the current work in this area. A Korean voice emotion database is used in [6] to demonstrate a two-stream-based emotion detection model based on convolutional neural networks (CNNs) and bidirectional long short-term memory (Bi-LSTM), with 90.38% and 94.91% performance attained. In order to identify speech emotion, 1D and 2D CNN LSTM networks were presented in [7]. The process of using log-mel spectrograms and raw audio samples to extract global contextual information and local correlations is examined. A SER system is proposed in [8] with an accuracy of 93.31% and 94.18% based on a 1-D deep convolutional neural network and various auditory characteristics. A DNN-based architectures for text-to-speech systems is suggested in [9], the result demonstrated the best performance, expressing sad and joyful emotions with over 60% of accuracy rate. A CNN based model for SER is used in [10] utilizing librosa package for extracting features and RAVDESS dataset, the model achieved an accuracy of 82.02%. A spoken emotion recognition system is constructed in [11] for WeChat that uses a random forest classifier, which has an 89% accuracy rate. SVM model is proposed in [12] to categorize the speech as belonging to one of the four emotions—happiness, fear, rage, or sadness—which ultimately resulted in higher accuracy rate for classification. A decision-trees methodology in conjunction with random forests method is suggested in [13] to represent the speech signals in order to categorize them. Recurrent neural networks with LSTM model for SER is proposed by [14], RAVDESS for datasets, the LSTM model achieved an accuracy rate of 78.2%, demonstrating it is robustness in SER classification. A technique for identifying speech emotions that combines silence removal with bidirectional LSTM and attention models is presented by [15]. By concentrating on emotive speech portions and eliminating unnecessary noise and stillness, it was able to increase accuracy. A dual-level model that uses LSTM and DS-LSTM architectures to combine MFCC and mel-spectrogram information, attaining cutting-edge accuracy for unimodal speech emotion recognition [16]. Transformer-based acoustic models (AMs) for hybrid speech recognition is presented in [17], and it achieves a 19% to 26% relative improvement over the best published hybrid result. Masked Predictive Coding is introduced in [18], an unsupervised pre-training technique that may be used for unsupervised pre-training with Transformer-based models. The experiment outperformed the best end-to-end model by more than 0.2%. Transformer-Transducer model is proposed in [19], the method achieved 6.37% of word error rates on the test clean test, and 15.30% on the other test set. A study conducted by [20] developed a hybrid LSTM-Transformer model that successfully captures contextual information and long-term dependencies for emotion recognition from speech audio recordings, the model recognition rate reached to 75.62%, 85.55% and 72.49%. In this paper, the hybrid model is tested on 3 benchmark datasets: TESS, SAVEE, and EMO-DB; however, it performed superiorly on the TESS database and obtained an accuracy of 99.64%, which marks a significant achievement in speech emotion recognition. The following Figure 2 shows

the basic architecture of the hybrid model, where the input layer feeds the data into both LSTM and Transformer layers, which are then merged together into a dense layer before delivering the final output.

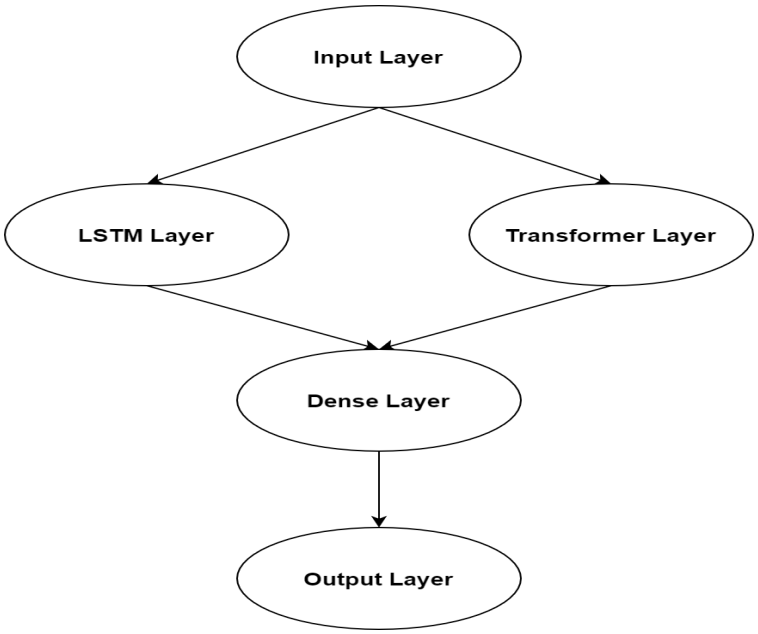

Figure 2. Basic architecture of LSTM + Transformer model

3. Proposed Methodology

In this section, we describe the techniques used to recognize speech emotions using audio processing methods and a pre-trained deep learning model. Firstly, datasets that contain .wav files are loaded with all different emotions. Secondly, audio signals are processed to extract meaningful features that are then fed into a deep learning model such as Transformer, Long Short-Term Memory (LSTM), and hybrid LSTM + Transformer methods that have been trained to classify emotions. The system's integration of cutting-edge machine learning techniques guarantees reliable performance in practical situations, Figure 3 below illustrates the overall process.

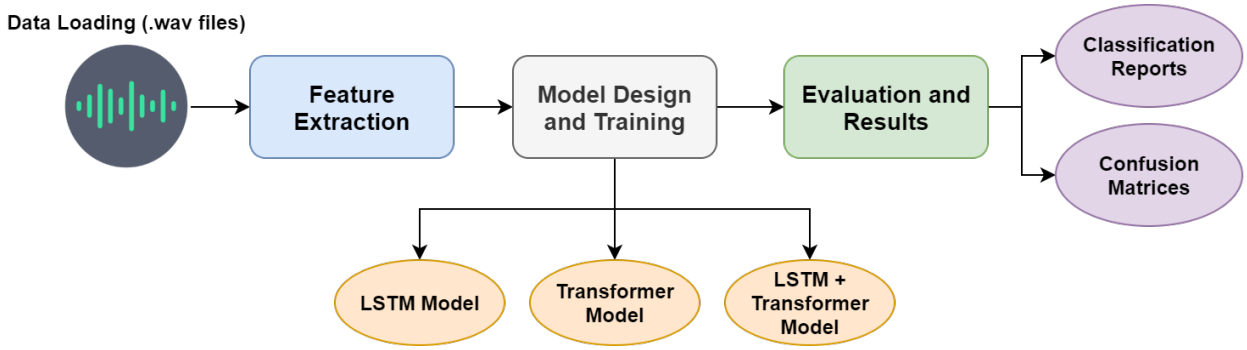

Figure 3. Proposed Framework

3.1. Data Loading

Audio files are loaded using datasets from three different sources which are: first Toronto Emotional Speech Set (TESS), the dataset includes recordings of 200 target sounds of old people, each of which is stated in one of seven emotional categories: angry, disgusted, afraid, happy, neutral, pleasant surprise, or sad. Secondly, EMO-DB (Berlin Emotional Speech Database) that contains 535 audio files of German speakers, and the structure of it is categories are as following: 'W': "anger", 'L': "boredom",'E': "disgust",'A': "fear", 'F': "happiness", 'T': "sadness", 'N': "neutral". Lastly, SAVEE (Surrey Audio-Visual Expressed Emotion) dataset it contains 480 acted emotions of audio and video recording from four male speakers with different emotions categories such as (anger, fear, disgust, sadness, happiness, surprise and neutral). Each speaker has 120 expressions which cover all emotion categories.

3.2. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) are used to preprocess raw audio data and create meaningful representations. It also widely used spectral features in speech processing [21]. The goal of this is to extract 40 MFCC features from every audio file using librosa.feature.mfcc(). Below is an example of computing Mel-Frequency Cepstral Coefficients (MFCCs):

MFCC shape: (40, time frames)
Average the MFCC values throughout different time frames:
Feature vector: mean(MFCC.T, axis=0) $\implies$ Shape: (40,)

3.3. Model Design and Training

In this section, the model's design and architecture will be discussed as shown in the following sections. The TESS, SAVEE, and EMO-DB datasets are then split into training data that contains 80% of the datasets and testing data that contains 20% of the datasets. A manual shuffling that contains an array of indices corresponding to the samples is used in both features (X) and labels (Y); it randomly shuffles the indices using np.random.shuffle(indices). This ensures that the data points are combined and mixed without changing how features and labels relate to one another. Each model is then trained with 50 epochs, a batch size of 32 to balance between convergence speed and performance stability. The Adam optimizer is used for its ability to adapt learning rates during training, sparse categorical cross-entropy loss is used for multi-class classification tasks where labels are integers rather than one-hot encoded vectors, and 40 features are represented by the input features' (1, 40) form to maintain a consistent structure for sequential model inputs.

3.3.1. LSTM model design: This model as shown in Figure 4, which uses LSTM, captures time-based patterns by running input sequences through two stacked LSTM layers (128 and 64 units, with tanh activation). To avoid overfitting, a dropout layer follows each LSTM layer. The last dense layer (7 units, softmax activation) is where the extracted features go through more changes after passing through a dense layer (32 units, ReLU activation). The output in the diagram should display the final model result, but it ties back to an "InputLayer," which isn't correct. This model was crafted to classify sequences or recognize emotions in speech.

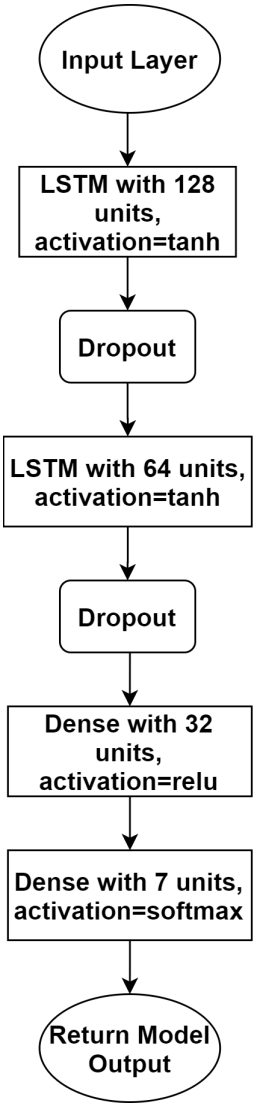

Figure 4. LSTM Model Architecture

3.3.2. ***Transformer model design:*** The diagram shown in Figure 5 illustrates a speech emotion recognition model based on Transformer architecture. It begins with an input layer shaped (1×40). After that several Multi-Head Attention layers pull out context-related connections. Each attention block has dropout, leftover links (Add), and Layer Normalization next to it to keep training steady. Dense layers with activation functions polish up how features are shown. Global Average Pooling boils down the features that were pulled out. The last dense layers give out a sorted result for figuring out emotions. This setup does a good job of catching time-based links and layered views from speech info.

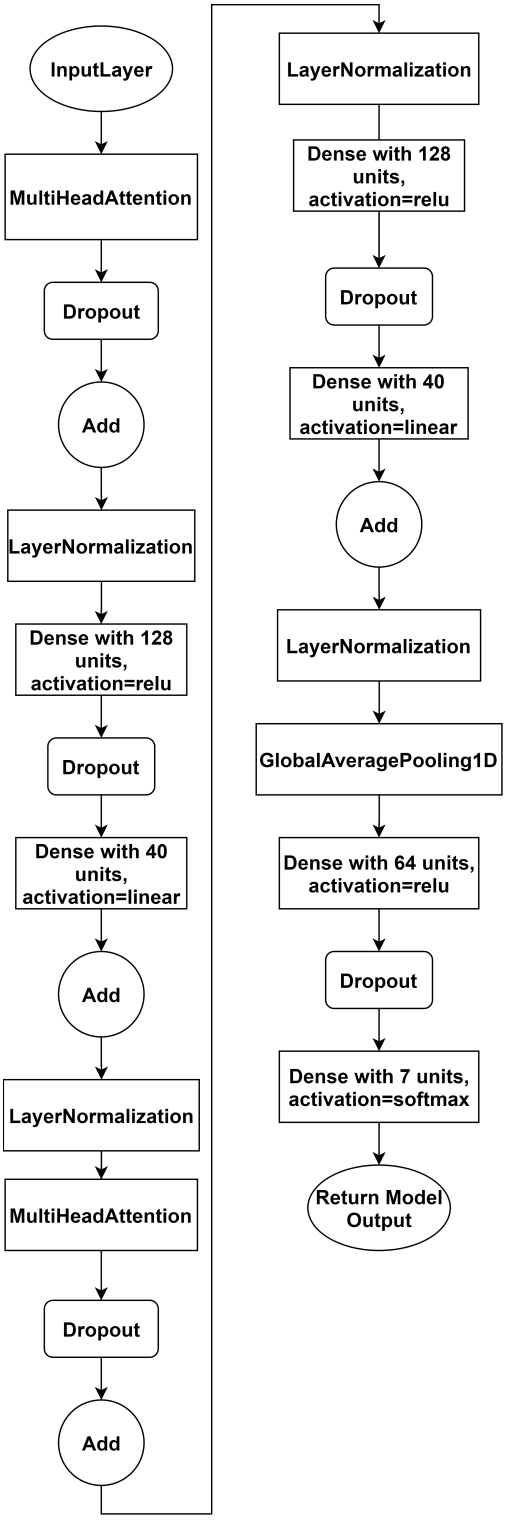

Figure 5. Transformer Model Architecture

3.3.3. ***Hybrid LSTM + Transformer model design:*** This combined LSTM and Transformer model illustrated in Figure 6 begins with an input layer. The input goes into an LSTM with 128 units and tanh activation. Next, it passes through a multi-head attention mechanism and dropout. The output then gets a dropout and an additive skip connection before layer normalization. A dense layer with 128 units and ReLU activation followed by another dropout, creates another skip connection. This leads to a final dense layer with 128 units and linear activation. The second branch uses two dense layers (32 units with ReLU and 7 units with softmax), dropout, layer normalization, and an LSTM with 64 units to make predictions. This setup uses Transformers' attention and LSTM's ability to remember sequences to spot emotions in speech.

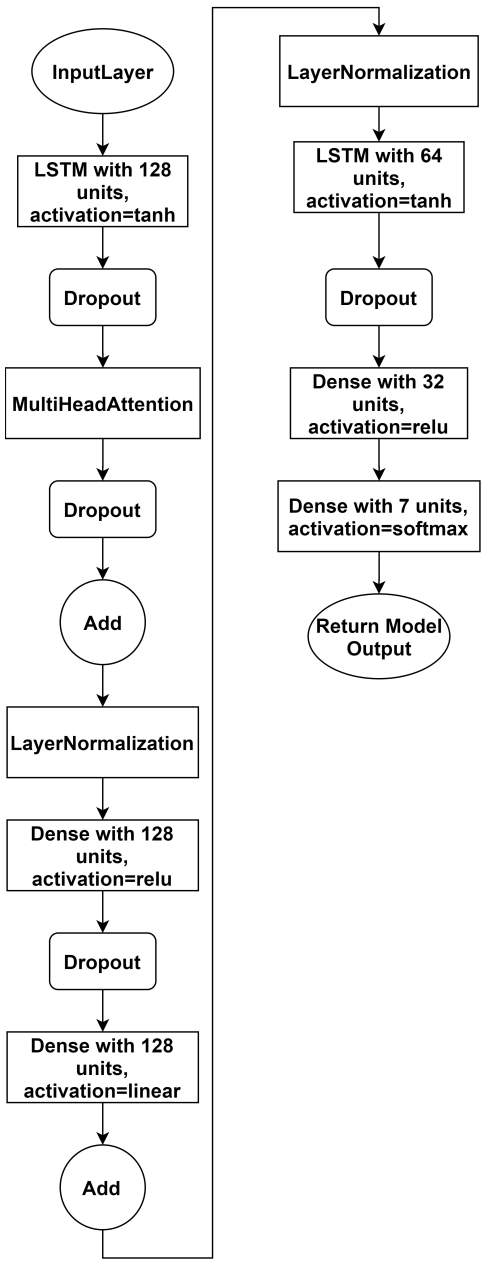

Figure 6. Hybrid LSTM + Transformer Model Architecture

3.4. Evaluation and Results

The test set is then used to assess the trained model's performance, which is gauged by a classification report such as precision, recall and f1-score for each class, in addition to confusion matrix to display the classification results for each class. Here are some examples of the formulas used for performance metrics:

Precision: Measures the accuracy of positive predictions of how many are correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Measures how many of the actual positive cases were predicted correctly.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: Measures the mean value of Precision and Recall, which finds the balance from both metrics.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where $TP =$ True Positive, $FP =$ False Positive, $FN =$ False Negative.

4. Experimental Results

In this section, we will examine the capability of LSTM, Transformer, and hybrid LSTM + Transformer models using performance measures obtained from confusion matrices and classification reports. The evaluation shows how well each model manages several classification task components, including precision, recall, and F1-score, as well as how well they reduce misclassifications. The specific findings and insights for each model will be covered in the sections that follow.

4.1. LSTM Model Analysis

LSTM model had an overall accuracy rate 97.50% on the TESS-DB. It performed well for anger, sadness, fear, disgust, and neutral emotions. These all had perfect precision, recall, and F1-measures. Pleasant surprise though, had lower recall (0.80), and happy had lower precision (0.84). For the EMO-DB, the accuracy fell to 73.83%. Anger and neutral did well (F1-measures of 0.82 and 0.94), but disgust and sadness had very poor results (0.33 and 0.52). The SAVEE-DB had the lowest accuracy at 65.62%. Results varied across emotions, with fear (F1-measure 0.31) and disgust (0.50) doing, but neutral performing quite well (0.79). The performance measures for this model are shown in the confusion matrices Figure 7 and classification reports Table 1 below.

Table 1. LSTM model classification report for TESS-DB, EMO-DB, and SAVEE-DB

TESS-DB					**EMO-DB**					**SAVEE-DB**				
Class	Precision	Recall	F1-score	Support	Class	Precision	Recall	F1-score	Support	Class	Precision	Recall	F1-score	Support
Fear	1.00	1.00	1.00	40	Anger	0.78	0.85	0.82	34	Anger	0.73	0.67	0.70	12
Pleasant_surprise	1.00	0.80	0.89	35	Boredom	0.67	0.67	0.67	12	Disgust	0.80	0.36	0.50	11
Sad	1.00	1.00	1.00	43	Disgust	0.33	0.33	0.33	3	Fear	0.40	0.25	0.31	8
Angry	1.00	1.00	1.00	42	Fear	0.62	0.67	0.64	12	Happiness	0.60	0.69	0.64	13
Disgust	1.00	1.00	1.00	37	Happiness	0.80	0.67	0.73	18	Neutral	0.70	0.92	0.79	25
Happy	0.84	1.00	0.92	38	Sadness	0.55	0.50	0.52	12	Sadness	0.71	0.71	0.71	14
Neutral	1.00	1.00	1.00	45	Neutral	0.94	0.94	0.94	16	Surprise	0.54	0.54	0.54	13

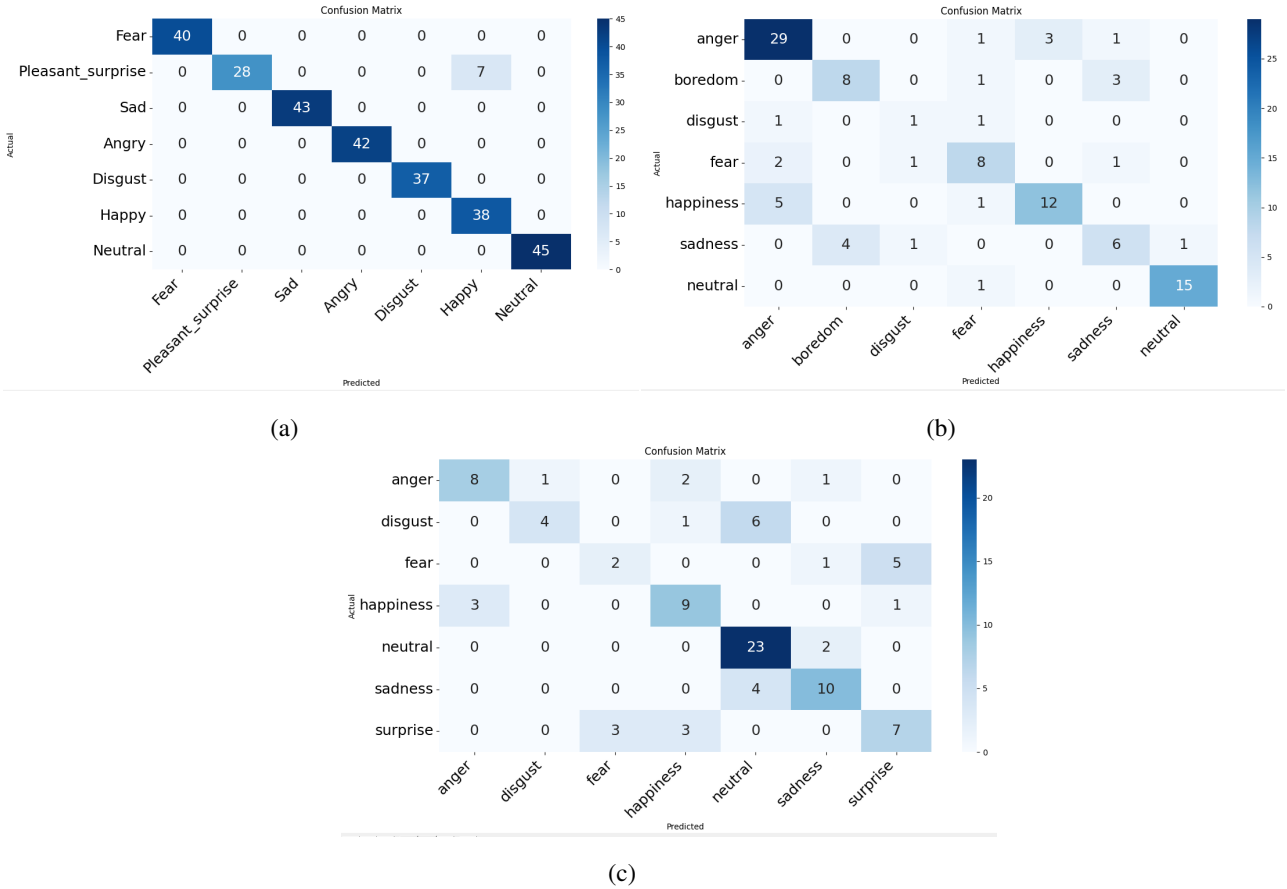

Figure 7. LSTM Model Confusion Matrix for (a) TESS-DB, (b) EMO-DB, and (c) SAVEE-DB

4.2. Transformer Model Analysis

The confusion matrices and classification reports described below in Figure 8 and Table 2 illustrate that the Transformer model performed almost perfectly in TESS-DB with accuracy of 98.21%, with all but pleasant surprises (F1-score: 0.93) and happy (0.94) having a perfect F1-score. Accuracy dropped to 68.22% in EMO-DB, with anger having a strong F1-score (0.84), but with weak results in disgust (0.33) and happiness (0.52). In SAVEE-DB the accuracy was lowest 58.33%, with a range of results in the various emotions. Specifically, neutral was strong (F1-score: 0.79), but fear (0.38) and happiness (0.47) were weak. Overall, the Transformer model was stronger than LSTM in the TESS-DB but weaker in EMO-DB and SAVEE-DB.

Table 2. Transformer model classification report for TESS-DB, EMO-DB, and SAVEE-DB

TESS-DB					**EMO-DB**					**SAVEE-DB**				
Class	Precision	Recall	F1-score	Support	Class	Precision	Recall	F1-score	Support	Class	Precision	Recall	F1-score	Support
Fear	1.00	1.00	1.00	40	Anger	0.78	0.91	0.84	34	Anger	0.60	0.50	0.55	12
Pleasant_surprise	0.97	0.89	0.93	35	Boredom	0.54	0.58	0.56	12	Disgust	1.00	0.36	0.53	11
Sad	1.00	1.00	1.00	43	Disgust	0.33	0.33	0.33	3	Fear	0.38	0.38	0.38	8
Angry	1.00	1.00	1.00	42	Fear	0.50	0.67	0.57	12	Happiness	0.41	0.54	0.47	13
Disgust	1.00	1.00	1.00	37	Happiness	0.78	0.39	0.52	18	Neutral	0.66	1.00	0.79	25
Happy	0.90	0.97	0.94	38	Sadness	0.86	0.75	0.80	16	Sadness	0.83	0.36	0.50	14
Neutral	1.00	1.00	1.00	45	Neutral	0.58	0.58	0.58	12	Surprise	0.46	0.46	0.46	13

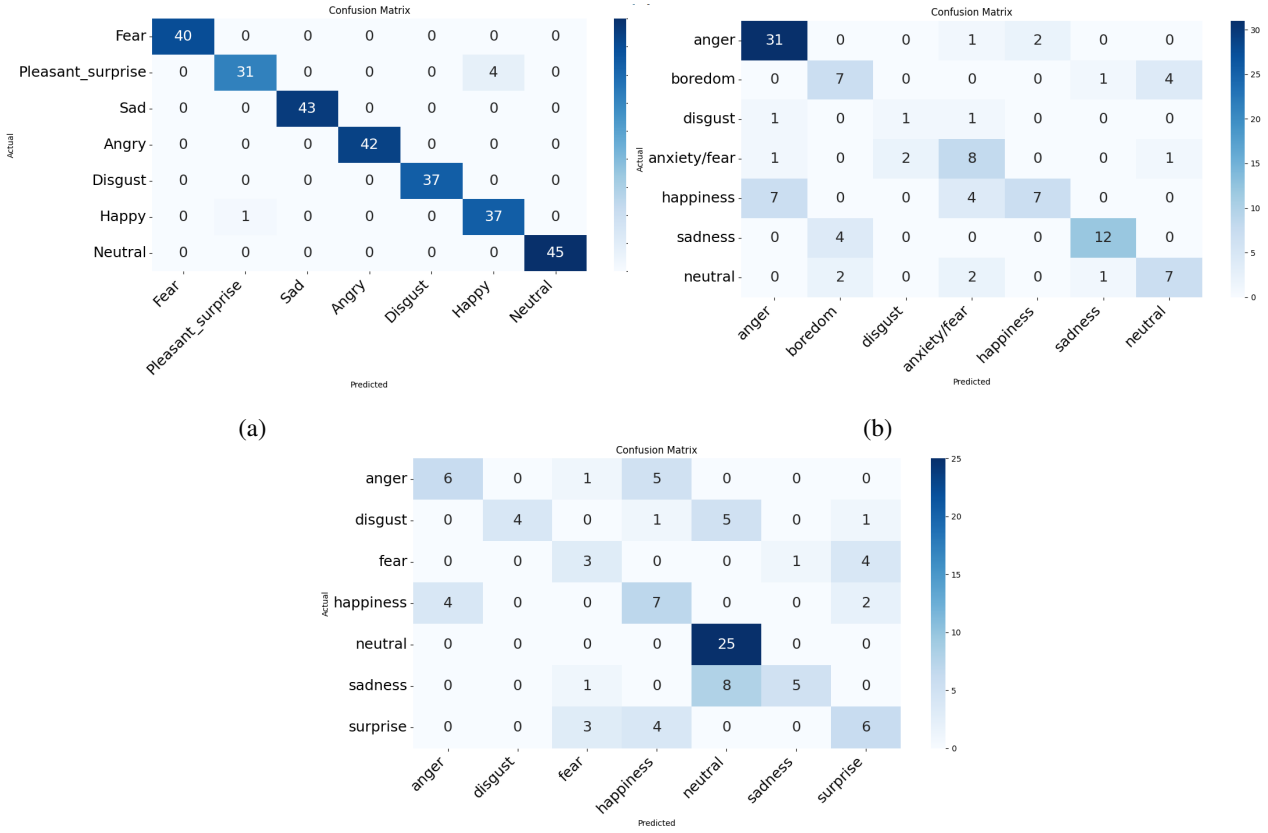

Figure 8. Transformer Model Confusion Matrix for (a) TESS-DB, (b) EMO-DB, and (c) SAVEE-DB

4.3. A Hybrid LSTM Transformer Model Analysis

The hybrid model showed an accuracy rate of 99.64% in the TESS-DB, outperforming the solo models in terms of performance for this specific dataset. It had perfect precision, recall, and F1-measures for all emotions except pleasant-surprise and happy, which both had an F1-measure of 0.99. In the EMO-DB, the model's accuracy has dropped to 62%. Sadness obtained the best performance with an F1-measure of 0.86, whereas disgust (0.25) and neutral (0.21) showed weak results. The SAVEE-DB the lowest accuracy at 56%. Here, neutral had the best F1-measure at 0.76, while disgust (0.17) and fear (0.29) performed weakly. The model excelled with TESS-DB but struggled with EMO-DB and SAVEE-DB, suggesting it lacks consistency across varied datasets perhaps because of the diversity and variability of emotional expressions in these kinds of datasets. Figure 9 and Table 3 below illustrate the confusion matrices and classification reports for the hybrid approach, presenting the performance measures.

Table 3. Hybrid model classification report for TESS-DB, EMO-DB, and SAVEE-DB

TESS-DB					**EMO-DB**					**SAVEE-DB**				
Class	Precision	Recall	F1-score	Support	Class	Precision	Recall	F1-score	Support	Class	Precision	Recall	F1-score	Support
Fear	1.00	1.00	1.00	37	Anger	0.70	0.91	0.79	34	Anger	0.53	0.67	0.59	12
Pleasant_surprise	0.98	1.00	0.99	50	Boredom	0.50	0.33	0.40	12	Disgust	1.00	0.09	0.17	11
Sad	1.00	1.00	1.00	42	Disgust	0.20	0.33	0.25	3	Fear	0.33	0.25	0.29	8
Angry	1.00	1.00	1.00	36	Fear	0.50	0.58	0.54	12	Happiness	0.43	0.46	0.44	13
Disgust	1.00	1.00	1.00	37	Happiness	0.62	0.28	0.38	18	Neutral	0.70	0.84	0.76	25
Happy	1.00	0.98	0.99	43	Sadness	0.76	1.00	0.86	16	Sadness	0.62	0.71	0.67	14
Neutral	1.00	1.00	1.00	35	Neutral	0.29	0.17	0.21	12	Surprise	0.43	0.46	0.44	13

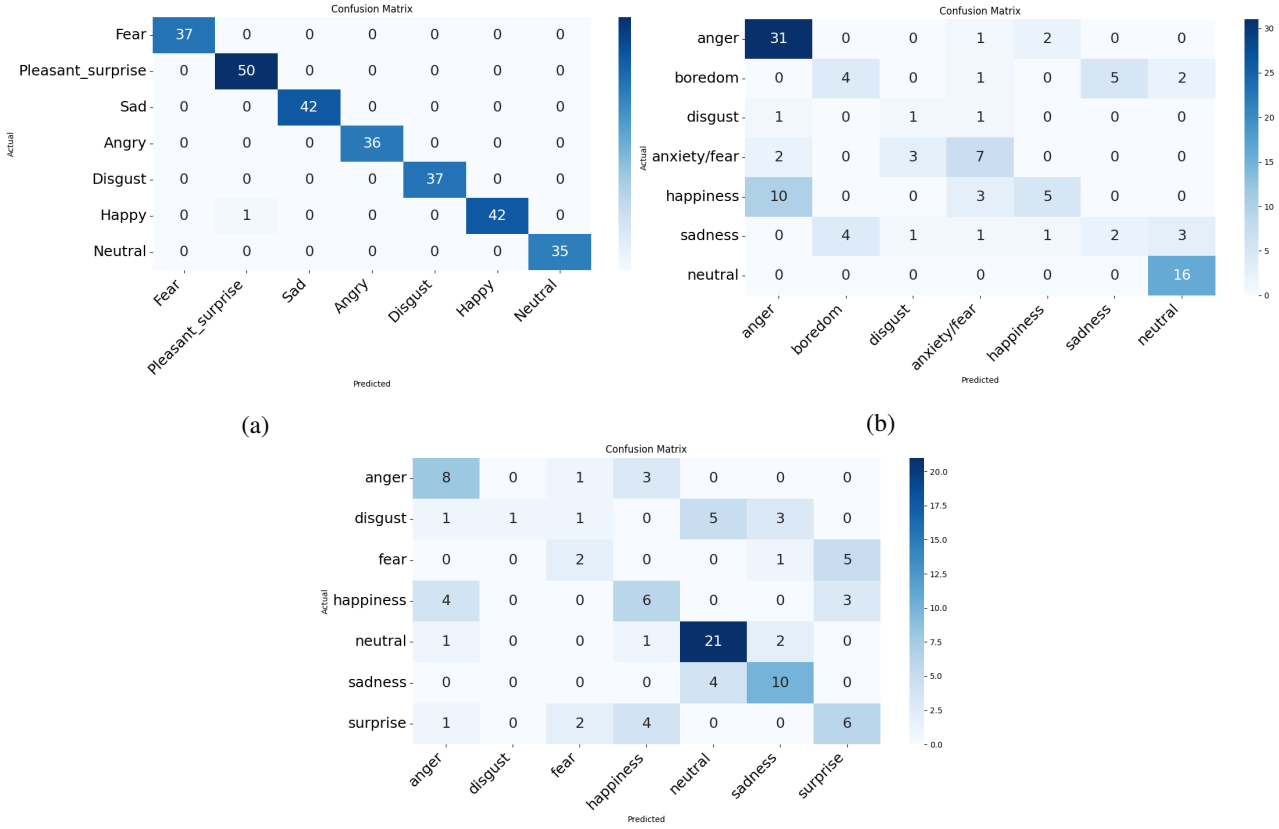

Figure 9. Hybrid Model Confusion Matrix for (a) TESS-DB, (b) EMO-DB, and (c) SAVEE-DB

5. Discussion

The hybrid LSTM + Transformer model performed outstandingly well on the TESS-DB, achieving 99.64% accuracy. This performance surpassed the standalone LSTM 97.50% and Transformer 98.21% models. The model's success comes from combining LSTM features to capture local time patterns such as speech signals with the transformer's ability to capture overall context through self-attention. TESS-DB is big, and well-organized data helps to fine-tune the hybrid, reducing overfitting and boosting how well it works in all aspects. However, in smaller datasets, the performance drops significantly. LSTM's basic structure handles short sequences better, while the hybrid's complexity adds unnecessary elements. Although the hybrid model offers theoretical advantages by combining sequence memory LSTM and global attention Transformer, its complex architecture may not be beneficial for smaller datasets such as EMO-DB and SAVEE-DB. These datasets may not provide enough diverse samples for the hybrid architecture to effectively generalize, causing the simpler solo LSTM to perform better. In contrast, in larger, well-balanced datasets, such as TESS-DB, the hybrid model demonstrates its superiority.

6. Conclusion

In this paper, an evaluation study of the performance of LSTM + Transformer versus standalone LSTM and Transformer for speech emotion recognition is conducted. Moreover, the study uses the three most used datasets, which are TESS, EMO-DB, and SAVEE, to test the proposed models. Furthermore, the outcome generated from

this study suggests that the hybrid model was somewhat outperformed by the solo Transformer and the solo LSTM models in the TESS dataset by 99.64%. For the EMO-DB and SAVEE-DB, the LSTM model had an edge over the hybrid and Transformer models in accuracy rate (78.83% - 65.62%). The hybrid model performs well when there is rich data; however, it struggles with smaller datasets, as demonstrated in the experimental result. Future work includes dynamic shuffling during training using lighter versions of Transformers and bidirectional LSTMs to handle longer dependencies. In addition, leveraging data augmentation methods and integrating attention mechanisms to enhance generalization and address overfitting challenges.

REFERENCES

1. T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features," *Sensors (Switzerland)*, vol. 20, no. 18, pp. 1–16, 2020.
2. F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
3. Y. Li, Y. Wang, X. Yang, and S. K. Im, "Speech Emotion Recognition Based on Graph-LSTM Neural Network," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, 2023.
4. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
5. I. Papastratis, "Speech Recognition: A Review of the Different Deep Learning Approaches." AI SUMMER, 2021.
6. A. H. Jo and K. C. Kwak, "Speech Emotion Recognition Based on Two-Stream Deep Learning Model Using Korean Audio Information," *Applied Sciences*, vol. 13, no. 4, 2023.
7. 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, 2019.
8. K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," *Electronics*, vol. 12, no. 4, 2023.
9. K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "Model Architectures to Extrapolate Emotional Expressions in DNN-Based Text-to-Speech," *Speech Communication*, vol. 126, pp. 35–43, 2021.
10. D. N. V. R. Reddy, "Speech Emotion Recognition Using Convolutional Neural Networks," *International Journal of Research in Applied Science and Engineering Technology*, vol. 12, no. 8, pp. 30–36, 2024.
11. S. Yan, L. Ye, S. Han, T. Han, Y. Li, and E. Alasaarela, "Speech Interactive Emotion Recognition System Based on Random Forest," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 1458–1462, 2020.
12. M. Jain *et al.*, "Speech Emotion Recognition using Support Vector Machine." Cornell University, 2020.
13. F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-Based Emotion Recognition Using Random Forests and Decision Tree," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 239–246, 2017.
14. Y. H. SaiDhruv, P. k, M. V. Vardhan, and J. S. J, "Speech Emotion Recognition Using LSTM Model," in *Proceedings of the International Conference on Recent Trends in Electronics and Communication*, pp. 692–697, 2023.
15. B. T. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, pp. 40–44, 2019.
16. J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech Emotion Recognition with Dual-Sequence LSTM Architecture," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6474–6478, 2020.
17. Y. Wang *et al.*, "Transformer-Based Acoustic Modeling for Hybrid Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6874–6878, 2020.
18. D. Jiang *et al.*, "Improving Transformer-based Speech Recognition Using Unsupervised Pre-training," *Cornell University*, 2019.
19. C.-C. Chiu *et al.*, "Transformer-Transducer: End-to-End Speech Recognition with Self-Attention," *Cornell University*, 2019.
20. F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022.
21. R. Zhao, X. Jiang, V. C. M. Leung, T. Wang, and S. Zhang, "Leveraging Cross-Attention Transformer and Multi-Feature Fusion for Cross-Linguistic Speech Emotion Recognition," *Cornell University*, 2025.