IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 13, 2022, accepted March 24, 2022, date of publication March 31, 2022, date of current version April 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3163856

Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files

FELICIA ANDAYANI[1], LAU BEE THENG[1], (Senior Member, IEEE), MARK TEEKIT TSUN[1], (Member, IEEE), AND CASLON CHUA[2]

[1]Faculty of Engineering, Computing, and Science, Swinburne University of Technology Sarawak Campus, Kuching, Sarawak 93350, Malaysia
[2]Faculty of Science, Engineering and Technology, Swinburne University of Technology, Melbourne, VIC 3122, Australia

Corresponding author: Felicia Andayani (fandayani@swinburne.edu.my)

This work was supported in part by the Swinburne University of Technology Sarawak Higher Degree by Research (HDR) Support Fund.

ABSTRACT Emotion is a vital component in daily human communication and it helps people understand each other. Emotion recognition plays a crucial role in developing human-computer interaction and computer-based speech emotion recognition. In a nutshell, Speech Emotion Recognition (SER) recognizes emotion signals transmitted through human speech or daily conversation where the emotions in a speech strongly depend on temporal information. Despite the fact that much existing research showed that a hybrid system performs better than traditional single classifiers used in SER, there are some limitations in each of them. As a result, this paper discussed a proposed hybrid Long Short-Term Memory (LSTM) Network and Transformer Encoder to learn the long-term dependencies in speech signals and classify emotions. Speech features are extracted with Mel Frequency Cepstral Coefficient (MFCC) and fed into the proposed hybrid LSTM-Transformer classifier. A range of performance evaluations was conducted on the proposed LSTM-Transformer model. The results indicate that it achieves a significant recognition improvement compared with existing models offered by other published works. The proposed hybrid model reached 75.62%, 85.55%, and 72.49% recognition success with the RAVDESS, Emo-DB, and language-independent datasets.

INDEX TERMS Attention mechanism, long short-term memory network, speech emotion recognition, transformer encoder.

I. INTRODUCTION

Emotion is an important aspect that exists in daily human activities. Emotions help people understand each other and assist people in decision-making [24], [26]. They also assist communication in the context of safety and security [24]. For example, when sharing with someone upset, we can be more careful and gentler to avoid hurting that person.

There are different modalities for recognizing human emotions, such as speech, text, and facial expressions. Speech is obviously an important channel and a source for studying human emotions [25]. SER is the task of recognizing emotions expressed through human speech. SER has played an important role in numerous applications, such as Human-Computer Interactions (HCI), Human-Robot Interfaces [1], intelligent call-centers [1], intelligent teaching systems [2], and many more. In addition, adding emotion recognition features is believed to be a crucial factor in creating a device that could act like a human [3]. Therefore, SER research is still actively pursued and has gained increasing interest from many researchers to develop a better performing recognition model.

Most SER research focuses on Machine Learning (ML) architectures to develop the SER model. This method involves extracting features from raw speech data. The extracted features are used as input to train the ML algorithm based on the samples of the input-output pairs. After the training, the ML algorithm predicts the emotions from the validation and testing data. Different types of features, such as prosodic, voice-quality, spectral, wavelet, spectrogram image, and deep features, have been widely used in current SER models. However, to date, no single feature set has been identified as a one-stop solution for recognizing emotion in speech data. Researchers often perform the testing or combine a vast number of features to gain some insight, and various feature selection methods can be used to remove

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

redundant features. The process of selecting the ML architecture used to perform the classification task is also crucial in SER, where the classification paradigm of the SER model must be able to process high-dimensional features at as low a computational cost as possible.

II. RELATED WORK

The emergence of Deep Learning (DL) has increased the efficiency possibilities for researchers to develop better performing SER. The models range from Deep Neural Network (DNN), Convolutional Neural Network (CNN), to Recurrent Neural Network (RNN) based applications.

In 2019, Lee *et al.* [4] used DNN to classify eight emotions on the Emo-DB dataset. They extracted 20 MFCCs with the delta and delta-delta values. They tested the model on four different sets of the extracted features. The highest results were obtained using 20 MFCCs with DNN, which achieved 69.4% recognition accuracy. The experiment also showed that DNN could perform better than a Support Vector Machine (SVM) model.

The successful application of CNN in image processing has motivated some researchers to develop end-to-end SER models. Such models often use CNN combined with LSTM to learn both spatial and temporal features. For instance, Tzirakis, Zhang, and Schuller [5] proposed an end-to-end model which used CNN and LSTM architectures. CNN was used to extract the features from the raw speech signals, while LSTM was used to learn the contextual information in the data and perform the final prediction. Their proposed model outperformed other state-of-the-art methods in terms of concordance correlation coefficient on the RECOLA dataset.

The use of the Attention Mechanism also shows possible improvements in the SER models. The Attention Mechanism is used to learn the critical features, and it is often combined with the LSTM architecture. Xie *et al.* [6] utilized the Attention Mechanism as an alternative to the forgetting gate in the LSTM architecture. In addition, they also applied the Attention Weighting Mechanism to distinguish the emotional saturation among the time segments. Besides, the Attention Weighting Mechanism could address the problem that different features may vary in their ability to distinguish emotions. The model used the frame-level speech features extracted using the OpenSmile library as an input. Their experiments on the CASIA, eNTERFACE, and GEMEP datasets showed recognition accuracy improvemen.

The successful application of the Attention Mechanism motivated Vaswasni *et al.* [7] to develop an architecture called the Transformer. The architecture was developed based on the Attention Mechanism, which allows parallelization and a global relationship between the input and output [7].The Transformer uses Multi-Head Attention Mechanisms instead of a single attention head. The model shows better performance in some natural language processing (NLP) tasks. Hence, some researchers have started to apply the Transformer architecture to emotion recognition tasks. However, researchers often combine different modalities to perform the job in the emotion recognition task. For example, Heusser *et al.* [8] proposed a reinforcement learning approachfor SER usinga pre-trained Transformer language model, which combined the SER task with the Speech Recognition and Text Emotion Recognition tasks. They used CNN-LSTM to perform the SER task anda pre-trained Transformer for the other two tasks. The proposed model was evaluated on the IEMOCAP database and achieved 73.5% and 71% recognition rates. Lee, Han, and Ko [9] also proposed a pre-trained Transformer model and CNN to perform the SER task. The model was trained and tested on the Emo-DB and IEMOCAP datasets. The recognition rates of the ‘‘speaker-dependent’’ and ‘‘speaker-independent’’ samples in the Emo-DB dataset were 94.23% Weighted Accuracy (WA) and 92.1% Unweighted Accuracy (UA) versus 88.43% WA and 86.04% UA, respectively. The IEMOCAP dataset’ recognition rates were 69.51% WA and 71.36% UA versus 66.47% WA and 67.12% UA, respectively. In [23], an improved model of Transformer was proposed and used inthe SER. They used different methods of positional encoding and Taylor Linear Attention (TLA) in Multi-Head Attention. Their model achieved 74.9% UA when tested on the Emo-DB dataset and 80% UAwhen tested on the URDU dataset.

Even though RNN has been widely used in SER research due to its ability to process sequential data and handle variable-length input, it suffers from a long-term dependency problem. However, the SER system requires a model that can sufficiently learn the long-term dependencies in the speech signal because emotions in speech signals strongly depend on temporal information.LSTM was developed to solve the RNN’ problem using its memory architecture. It can remember the information for an extended period of time. Nevertheless, LSTM might not work well on longer-term dependencies [15].

The Transformer, a Multi-Head Attention Mechanism, was introduced in 2017 [7], and it has been widely used in many NLP fields. In contrast to RNN and LSTM, it can be parallelized. Moreover, it has shown outstanding performance in broad NLP tasks. However, the Transformer’s weakness is that it loses the sequential information of its position and needs to re-compute the entire history in the context window at each time step [11].

Looking at the advantages and disadvantages of the LSTM and Transformer architectures, we have gained insight into combining them to enjoy the benefits of both architectures while preventing their respective drawbacks. Moreover, both architectures have been incorporated into different research fields and achieved state-of-the-art results, such as language modeling [15], [19], text generation [10], and modeling multi-leg trips [11].

However, the combination has not been used in SER. Thus, the primary goal of this study is to combine recent advances in the LSTM and Transformer architectures in the SER system and investigate the impact of the combination on improving the SER classification performance. The state-of-the-art results in [10], [11] have motivated us to explore the hybrid

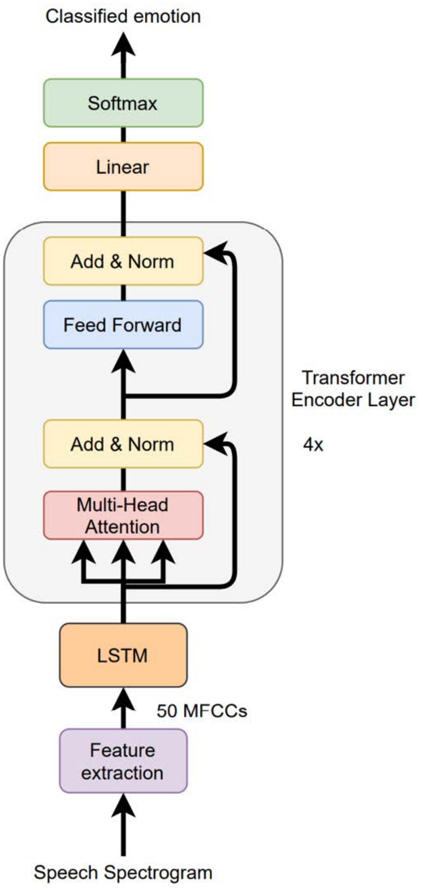

FIGURE 1. Overview of the proposed LSTM-Transformer model.

LSTM-Transformer model to solve the emotion classification problems in SER.

We propose a hybrid LSTM-Transformer architecture that improves the SER classification performance via learning the long-term dependencies. In addition, we investigated the performance of the LSTM-Transformer model on recognizing emotions in speech signals by learning the long-term dependencies, which have not been explored previously.The proposed LSTM-Transformer model replaces the positional encoding in the Transformer architecture with the LSTM recurrent process to learn the hidden state of the input features. In addition, instead of using a single Attention layer on the LSTM, we combined the LSTM with the Multi-Head Attention Mechanism in the Transformer encoder layer.

The maincontributions of this work are the design, development, and evaluation of the model, selection of parameters, and selection of the MFCC feature for th LSTM-Transformer model to classify emotions. This work utilizes the RAVDESS, Emo-DB, and language-independent datasets. Our language-independent dataset is developed froma combination of RAVDESS and Emo-DB datasets.

III. PROPOSED SOLUTION

The overview of the proposed approach is depicted in Fig. 1, where the speech samples are pre-processed and converted into spectrograms. It is then followed by the MFCC features being extracted from the spectrograms and the mean statistics from the extracted MFCC features are calculated to compute the feature vector. Upon extracting the features, standardization is applied to have standardized distributed data. Towards the end of the pipeline, the feature vector is used for training and testing the LSTM-Transformer model.

A. PRE-PROCESSING AND FEATURE EXTRACTION

Diverse datasets have different characteristics, with some containing noise while others are clean recordings. Some datasets also contain silenceat the beginning or end of the recordings, and the recorded speech data durationmay vary in some datasets. Therefore, speech data requires pre-processing to maintain consistent training and testing data.

The samples from the selected datasets were loaded and resampled to 22050 kHz so that the language-independent dataset could adhere to a consistent sampling rate. Then, the silence parts at the beginning and the end were removed, and the signal containing the speech information was obtained for furtherprocessing.

Upon completin the pre-processing, the speech samples were converted into spectrograms as the input to the proposed model. This study extracted the Mel Frequency Cepstral Coefficient features from the spectrograms [35], one of the most widely used audio features in speech processing applications. The MFCC is often used due to its ability to mimic the human hearing system and provide information on the human vocal tract' shape [35]. The feature extraction process is implemented using the Librosa library [12]. To obtain the spectrogram, the Short-Term Fourier Transform (STFT) was performed on the speech signal with a window length of 1024 and a hop length of 512. The MFCCs were obtained by applying Mel filters, taking the log-magnitude, and applying a Discrete Cosine Transform (DCT) to the spectrograms.

In this research, the mean of 50 MFCCs was obtained and used for further training and testing the proposed model. According to [27], emotional characteristics were inherited from the whole speech file and were not affected by the details in the individual frames. Thus, the mean values of extracted MFCCs were calculated and mapped into the feature vector to avoid losing the temporal information when fed to static classifiers, such as Neural Networks. Besides, the mean values of MFCCs were calculated to have a fixed-length feature vector [28]. Such features have been widely used in SER systems and have achieved state-of-the-art results [21], [22], [29]–[32].

Lastly, standardization was applied before the recognition step to have standardized distributed data. In this research, Standard Scaling was used and computed using the *Scikit-learn* library to standardize the features by removing the mean and scaling to unit variance [20].

B. LSTM-TRANSFORMER MODEL

We combined both the LSTM and the Transformer layers to learn the long-term dependencies in speech signalsfor emotion recognition.

In this study, the LSTM layer replaces the positional encoding in the Transformer architecture. In addition, the use of positional encoding in the Transformer architecture is believed to be the source of higher computational costs because it must learn the entire history from the beginning at each time step [11]. With the recurrent process of the LSTM, the hidden state of the input features is preserved.

Moreover, LSTM has been developed to solve the short-term memory proble; however, the LSTM alone is insufficient to solve the longer-termdependency problem [15]. Therefore, the Multi-Head Attention in the Transformer encoder layer has been modeled and integrated to improve the model's ability to learn the long-term dependencies. Multi-Head Attention can jointly attend to information from the extracted features at different sequence positions. The Transformer encoder layer also contains a feed-forward network layer with ReLu activation and layer normalization.

The Transformer layer allows for parallelization, which LSTM cannot. This layer enables the model to perform faster in learning the long-term dependencies. The combination of both architectures is expected to be better at understanding the long-term dependencies. Thus, it is suitable for the SER model because the emotion in speech depends highly on temporal information. Finally, the Linear and Softmax layers are applied to get the final predicted emotion. Dropouts are also applied to the LSTM, the Transformer encoder, and the final Linear layer to avoid overfitting.

IV. EVALUATION

This section discusses the specifications of datasets used to evaluate the model's performance and the experimental settings.

A. DATASETS

The performance of the proposed model was evaluated on three dataset: RAVDESS, Emo-DB, and the language-independent dataset. The datasets were chosen based on their credibility and wide usage in most research in the SER field. These widely-used datasets provide efficiency in comparing the performance of the proposed model with other studies utilizing the same datasets. Table 1 shows the overview of datasets involved in this experiment.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [13] is a publicly available dataset containing recorded speech and song data. However, in this research, only the recorded speech data were used.The speech data were recorded in American English by 24 actors (12 males and 12 females). The speech data were recordedat normal and strong intensities, except for the neutral emotion. One thousand four hundred forty samples were obtained from the speech data, containing eight different emotions: happiness, sadness, anger, fear, surprise, disgust, calm, neutral.

Another dataset used was the Emo-DB [14]. It contains 535 speech samples in German, recorded by ten actors five males and five females). Moreover, the dataset covers seven different emotions: anger, fear, disgust, boredom, neutral, happiness, an sadness. It is also a publicly available dataset in the SER research field.

TABLE 1. Overview of the datasets used.

Emotion	RAVDESS	EMO-DB	LANGUAGE-INDEPENDENT
Anger	192	127	319
Happiness	192	71	263
Sadness	192	62	254
Fear	192	69	261
Disgust	192	46	238
Neutral	96	79	175
Boredom	-	81	-
Surprised	192	-	-
Calm	192	-	-
Total	1440	535	1510

As there exist many spoken languages around the world, developing a versatile SER model that is not limited to a specific language is important. Therefore, we analyze the language-independent dataset, which contains speech from English and German languages, to develop a SER model that is oblivious to the language of the spoken speech. As a result, we could achieve a versatile SER model that will broaden the application of the proposed SER model that is not limited to a specific language. The language-independent dataset was developed by combining the RAVDESS and Emo-DB datasets. However, only the emotion components contained in both datasets were used.

Furthermore, investigating the model in language-dependent and language-independent datasets provides some insights into how the emotion patterns are shared across different languages.

B. EXPERIMENTAL SETTINGS

The experiments were conducted on Google Colaborator equipped with an Intel(R) Xeon(R) CPU @ 2.20GHz, 25GB RAM, and NVIDIA Tesla T4 GPUs. We also used the Python programming framework to implement the proposed model.

The model used 64 dimensions of the LSTM hidden layer and four layers of the Transformer encoder layer. Each Transformer encoder layer consisted of four heads of self-attention followed by 512 dimensions of feed-forward layer and layer normalization. The Linear and Softmax layers further processed the final output to predict the emotion.

The batch size was set to 60 because it showed the best performance, while the SGD optimizer was adapted with a learning rate of 0.001 and 0.9 momentum during the training. The model was trained for 750 epochs. Lastly, 10-fold cross-validation [21], [22], [31] and a hold-out procedure were used to assess the model's performance. Both procedures are common techniques used in evaluatingSER models performance. In addition, 10% of the samples from each emotion on each dataset were held-out for final testing. The rest of the samples were used to perform the 10-fold cross-validation.

TABLE 2. The results of 10-fold cross-validation on three datasets.

Fold	RAVDESS	EMO-DB	LANGUAGE-INDEPENDENT
1	69.77%	72.92%	72.79%
2	65.12%	91.67%	68.38%
3	69.77%	83.33%	71.32%
4	70.54%	79.17%	71.32%
5	66.67%	75.00%	70.59%
6	72.87%	75.00%	73.53%
7	69.77%	72.92%	77.04%
8	68.22%	72.92%	67.41%
9	68.99%	72.34%	60.74%
10	74.42%	76.60%	68.89%
Mean	**69.61%**	**77.19%**	**70.20%**

TABLE 3. Performance of the proposed model on the hold-out test set on three datasets.

Datasets	Weighted Accuracy (WA)	Unweighted Accuracy (UA)
RAVDESS	77.33%	75.62%
Emo-DB	87.72%	85.55%
Language-Independent	71.43%	72.50%

The 10-fold cross-validation was performed as emotion-independent cross-validation. The weighted accuracy (WA) and unweighted accuracy (UA) were calculated to evaluate the performance of the proposed model. A comparative study was also conducted to analyze the proposed model's performance.

V. RESULTS AND DISCUSSIONS

The model was first evaluated using 10-fold cross-validation on each dataset in the experiments.Ten percent of each emotion on each dataset was held out for testing the model after the 10-fold cross-validation process. The rest of the samples were shuffled and randomly split into ten folds of approximately equal size. Each k subset was used for validation, and the remaining k-1 subsets were used for training the model. The process was repeated ten times. The accuracy results of 10 folds were obtained, and the average accuracy of the ten folds was used to determine the performance. Table 2 shows the results obtained from the 10-fold cross-validation calculated from the validation set.

Table 2 shows that the highest recognition was obtained when the model was trained with the Emo-DB dataset, achieving an average of 77.19% recognition rate. The model also acquired 69.61and 70.20% average recognition rates on the RAVDESS and language-independent datasets. Although RAVDESS achieved the lowest average recognition rate, the model achieved up to 74.42% recognition rate and performed well on the hold-out test set. Besides, a 70.20% recognition rate with the language-independent dataset showed that the model could predict emotions in different languages.

Subsequently, the best model for each dataset was selected and used to plot the accuracy and loss curves. Fig. 2, Fig. 3, and Fig. 4 show the learning curves on RAVDESS, Emo-DB, and language-independent datasets, respectively. The learning curves show that the validation loss and the validation accuracy improve with the number of iterations. However, they become rather constant after around 300 epochs. Nevertheless, the model achieved better performance when trained on 750 epochs instead of 300 epochs, which means increasing the number of iterations can improve the model's performance, despite the minor improvements in validation loss and accuracy. It is similar to the study done by [33]. In [33], the model performed better on 4000 epochs than on 100 epochs.

Moreover, the validation accuracycurve of the Emo-DB dataset, as shown in Fig. 3, achieved 91.67% accuracy, which is higher than the training accuracy curve. It may be caused by the imbalanced division of the dataset on that particular fold and the complexity of the validation set, which requires further investigation.The validation loss curves on the RAVDESS and language-independent datasets can be further investigate, as they seem to be relatively high in the learning curves.

The selected best model for each dataset was tested on the 10% hold-out test set. The WA and UA were calculated from the test set, as shown in Table 3. According to Table 3, the best accuracy was on the Emo-DB dataset, followed by the RAVDESS and the language-independent datasets. In addition, the RAVDESS result (77.33% WA and 75.62% UA) was slightly lower than the result obtained from the Emo-DB dataset. It is an acceptable rate due to the complexity of the RAVDESS dataset. Therefore, it requires further preprocessing steps than the Emo-DB dataset, where the samples contain less noise. Consequently, when tested on the Emo-DB dataset, the model achieved a higher recognition rate (87.72% WA and 85.55% UA). Moreover, when tested on the language-independent dataset, the model achieved a 71.43% WA and a 72.50% UA The UA performed in a language-independent dataset indicates that the model can deal with imbalanced data among emotional classes in that dataset.

Furthermore, the results obtained from the hold-out test set were higher than the average accuracy obtained from the 10-fold cross-validation,indicating that the model could generalize well on the new data while not demonstrating any overfitting.

The confusion matrix for the best classification result on each dataset was generated to show the actual predicted emotions. Each row in the confusion matrix shows the actual emotion, whereas each column in the confusion matrix shows the predicted emotion. The diagonal values indicate the number of correctly predicted emotions.

The confusion matrices on RAVDESS, Emo-DB, and language-independent datasets are shown in Table 4, 5, and 6, respectively. Table 4 (RAVDESS) shows that the best

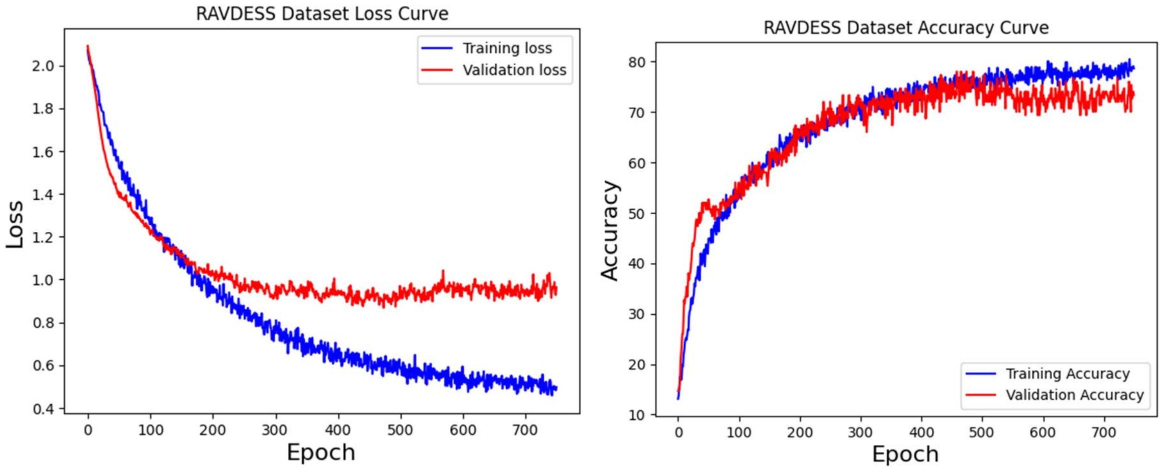

FIGURE 2. RAVDESS dataset loss (left) and accuracy (right) curves from fold 10.

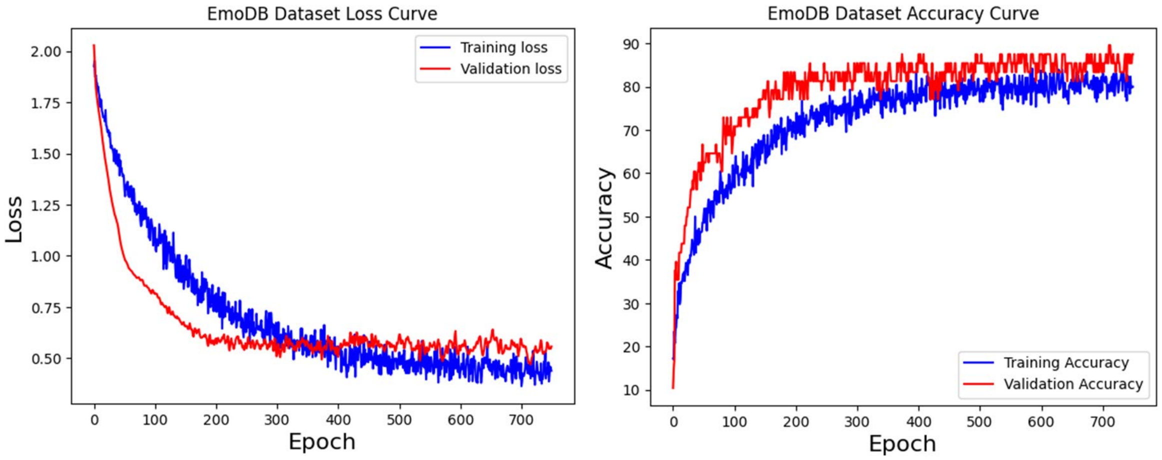

FIGURE 3. EmoDB dataset loss (left) and accuracy (right) curve from fold 2.

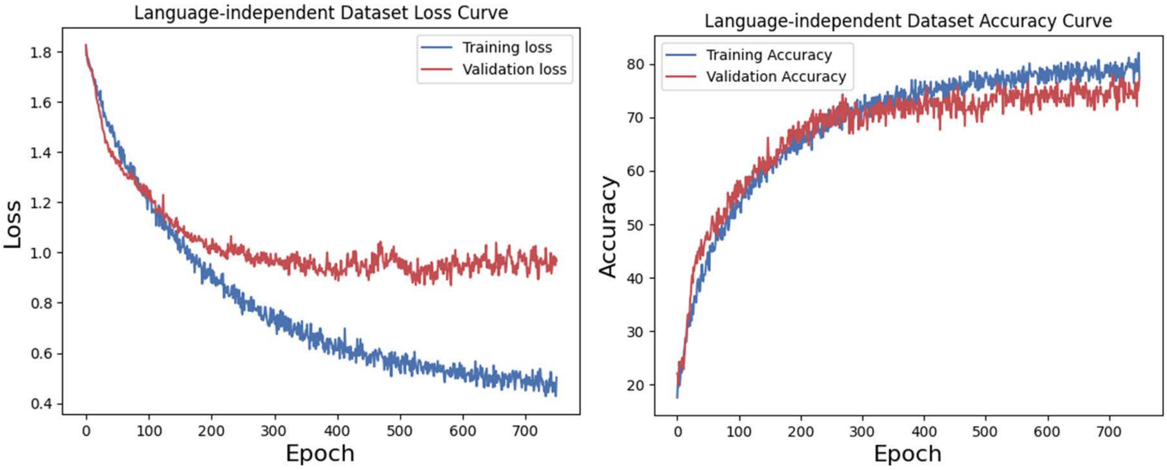

FIGURE 4. Language-independent dataset loss (left) and accuracy (right) curve from fold 7.

recognition was achieved for the calm emotion (95%). Our model acquired the highest recognition for the calm emotion among the models proposed in other studies to the best of our knowledge. The neutral emotion achieved the least

TABLE 4. **Confusion matrix of the proposed model on RAVDESS with 75.62% of unweighted accuracy. The model achieved the best recognition for calm emotion.**

	Predicted Emotion							
Actual Emotion	Anger	Happiness	Sadness	Fear	Disgust	Neutral	Surprised	Calm
Anger	17 85%	0 0%	0 0%	1 5%	1 5%	0 0%	1 5%	0 0%
Happiness	1 5%	13 65%	0 0%	2 10%	1 5%	0 0%	3 15%	0 0%
Sadness	0 0%	0 0%	18 90%	0 0%	1 5%	1 5%	0 0%	0 0%
Fear	0 0%	1 5%	4 20%	13 65%	0 0%	0 0%	2 10%	0 0%
Disgust	1 5%	2 10%	1 5%	0 0%	16 80%	0 0%	0 0%	0 0%
Neutral	0 0%	1 10%	1 10%	0 0%	0 0%	5 50%	2 20%	1 10%
Surprised	2 10%	1 5%	0 0%	2 10%	0 0%	0 0%	15 75%	0 0%
Calm	0 0%	0 0%	1 5%	0 0%	0 0%	0 0%	0 0%	19 95%

TABLE 5. **Confusion matrix of the proposed model on Emo-DB with 85.55% of unweighted accuracy. The model achieved the best recognition for anger, fear, and sad emotions.**

	Predicted Emotion						
Actual Emotion	Anger	Happiness	Sadness	Fear	Disgust	Neutral	Boredom
Anger	13 100%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
Happiness	2 25%	6 75%	0 0%	0 0%	0 0%	0 0%	0 0%
Sadness	0 0%	0 0%	7 100%	0 0%	0 0%	0 0%	0 0%
Fear	0 0%	0 0%	0 0%	7 100%	0 0%	0 0%	0 0%
Disgust	1 20%	1 20%	0 0%	0 0%	3 60%	0 0%	0 0%
Neutral	0 0%	0 0%	0 0%	1 12.50%	0 0%	6 75%	1 12.50%
Boredom	0 0%	0 0%	0 0%	0 0%	0 0%	1 11%	8 89%

TABLE 6. **Confusion matrix of the proposed model on language-independent dataset with 72.50% of unweighted accuracy. The model achieved the best recognition for neutral emotion.**

	Predicted Emotion					
Actual Emotion	Anger	Happiness	Sadness	Fear	Disgust	Neutral
Anger	25 78%	1 3.10%	1 3.10%	2 6.20%	2 6.20%	1 3.10%
Happiness	3 11%	15 56%	2 7.40%	4 15%	1 3.70%	2 7.40%
Sadness	0 0%	1 3.80%	19 73%	1 3.80%	4 15%	1 3.80%
Fear	3 11%	2 7.40%	2 7.40%	17 63%	0 0%	3 11%
Disgust	1 4.20%	1 4.20%	2 8.30%	1 4.20%	17 71%	2 8.30%
Neutral	0 0%	0 0%	0 0%	0 0%	1 5.60%	17 94%

recognition rate, which is affected by the small sample size of neutral emotion in the RAVDESS dataset. Table 5 (Emo-DB) shows that the model obtained the best recognition results for anger, fear, and sadness, which achieved a 100% recognition

TABLE 7. Performance of the proposed model against other publications on the RAVDESS dataset.

Authors	Features Used	Model	Unweighted Accuracy (UA)
Parry et al. [16]	MFCC	LSTM	53.97 %
Jalal et al. [17]	Fundamental frequency (F0), MFCC, delta & delta-delta log-energy	BLSTM and Capsule routing	56.20%
Zeng et al. [18]	Spectrogram	Gated ResNets	60.35%
This paper	**MFCC**	**LSTM-Transformer**	**75.62%**

TABLE 8. Performance of the proposed model against other publications on the Emo-DB dataset.

Authors	Features Used	Model	Unweighted Accuracy (UA)
Parry et al. [16]	MFCC	LSTM	59.67%
Kerkeni et al. [21]	Modulation Spectral Feature (MSF) & MFCC	LSTM	69.55%
Jing, Manting, and Li [23]	Log-Mel Filterbank Energies (LFBE)	Transformer	74.9%
Kerkeni et al. [22]	Modulation Spectral Feature (MSF) & MFCC	LSTM	83.00%
This paper	**MFCC**	**LSTM-Transformer**	**85.55%**

rate. It reached a lower recognition for disgust emotion mainly due to the small training sample size. On the other hand, the model performed well for all the emotions in the language-independent dataset, except for happiness and fear, as shown in Table 6. In addition, there is an improvement in the neutral emotion recognition when evaluated under a language-independent method, indicating that the neutral emotion shares the same emotion pattern between the English and German languages. However, the low recognition of happiness and fear emotions shows that both emotions may have different patterns between the English and German languages, which requires further investigation to differentiate them. Nevertheless, the model performed well on the language-independent dataset. Thus, it will perform well when implemented in real-time applications where a versatile SER model is required.

The results obtained in the experiments are compared with other LSTM and Transformer models proposed by existing researchers. The comparative study was done as they shared the same datasets. The comparisons should be viewed as indicative benchmarking instead of absolute one-to-one performance rankings [34] because the methods used between researchers are different, such as the experimental settings, the features utilized, and the classification models. Table 7 and Table 8 show the performance comparison of the proposed model with other models on the RAVDESS and Emo-DB datasets, respectively.

Our proposed hybrid model outperformed the models of Parry *et al.* [16], Jalal *et al.* [17], and Zeng *et al.* [18] using the RAVDESS dataset in our testing. Likewise, our proposed model also performed better than Parry *et al.* [16], Kerkeni *et al.* [21], Kerkeni *et al.* [22], and Jing, Manting, and Li [23] when evaluated using the Emo-DB dataset. In addition, our proposed model outperformed Jing, Manting, and Li's [23] Transformer model with positional encoding, which demonstrated that incorporation of LSTM into the Transformer model improved the recognition performance by maintaining the hidden state of the input features with long-term dependency.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a hybrid model for SER by combining the LSTM and Transformer architectures. The strengths of both architectures are adapted to improve the recognition performance in the SER. Our hybrid model performed better at learning the long-term dependencies in speech signals by preserving the hidden state of input features using LSTM and the use of Multi-Head Attention on the Transformer encoder layer and the MFCC feature vectors. The proposed hybrid model is able to learn the temporal information from the frequency distributions in the MFCCs of each emotion in both language-independent and language-dependent datasets. The hybrid model's effectiveness is shown through the results obtained from the experiments in this research. The proposed model was evaluated on the RAVDESS, Emo-DB, and language-independent datasets. The recognition rate improved from 15.27% to 21.65% for the RAVDESS dataset compared with other published models [16]–[18] and improved from 2.55% to 25.88% for the Emo-DB dataset compared with other published works [16], [21]–[23]. In addition, the proposed model achieved a WA of 71.43% and a UA of 72.50% on the language-independent dataset.

In conclusion, the proposed model shows a significant improvement for the language-dependent datasets, RAVDESS and Emo-DB. Besides, the results from the language-independent dataset indicate that the model can perform well in a mixed language situation. It also shows that the happiness and fear emotions have different patterns between languages, whereas the neutral emotion shows a similar pattern.

In the future, an improvement to the pre-processing method may be carried out, especially on the language-independent dataset. It was discovered that resampling the speech data on the language-independent dataset could affect the quality of the speech samples, which may have deteriorated the recognition outcomes. Besides, our proposed hybrid model can be improved by incorporating additional feature types in the SER training and recognition, such as prosodic or deep features. It can also be enhanced for real-time SER systems by incorporating the variable input sequence and raw audio inputs. Data augmentation can be applied to overcome the problem of data shortage in training and testing datasets. Datasets from other languages can be added to improve the language-independent emotion recognition ability. The proposed hybrid model on cross-corpus speech emotion recognition should also be investigated.

REFERENCES

[1] J. Sidorova, "Speech emotion recognition," Master Thesis, Univ. Pompeu Fabra, Barcelona, Tech. Rep., 2007, doi: 10.13140/RG.2.1.3498.0724

[2] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "A review on speech emotion recognition: Case of pedagogical interaction in classroom," in *Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Fez, Morocco, May 2017, pp. 1–7, doi: 10.1109/ATSIP.2017.8075575.

[3] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers Comput. Sci.*, vol. 2, p. 14, May 2020, doi: 10.3389/fcomp.2020.00014.

[4] K. H. Lee, H. Kyun Choi, B. T. Jang, and D. H. Kim, "A study on speech emotion recognition using a deep neural network," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Jeju Island, South Korea, Oct. 2019, pp. 1162–1165, doi: 10.1109/ICTC46691.2019.8939830.

[5] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5089–5093, doi: 10.1109/ICASSP.2018.8462677.

[6] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 11, pp. 1675–1685, Jul. 2019, doi: 10.1109/TASLP.2019.2925934.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[8] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," 2019, *arXiv:1912.02610*.

[9] S. Lee, D. K. Han, and H. Ko, "Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition," *Sensors*, vol. 20, no. 22, p. 6688, Nov. 2020, doi: 10.3390/s20226688.

[10] A. Tanikawa. (2020). *Text Generation With LSTM+Transformer Model (Japanese)*. [Online]. Available: https://note.com/diatonic_codes/n/nab29c78bbf2e

[11] Y. Sakatani, "Combining RNN with transformer for modeling multi-leg trips," in *Proc. ACM WSDM WebTour*, Jerusalem, Israel, 2021, pp. 50–52.

[12] B. McFee *et al.*, "Librosa/librosa," Zenodo, 2020, doi: 10.5281/zenodo.591533.

[13] 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, doi: 10.1371/journal.Pone.0196391.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1517–1520, doi: 10.21437/Interspeech.2005-446.

[15] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 2978–2988. [Online]. Available: https://aclanthology.org/P19-1285.pdf

[16] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1650–1656, doi: 10.21437/Interspeech.2019-2753.

[17] A. Md Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1701–1705, doi: 10.21437/Interspeech.2019-3068.

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

[19] Z. Huang, P. Xu, D. Liang, A. Mishra, and B. Xiang, "TRANS-BLSTM: Transformer with bidirectional LSTM for language understanding," 2020, *arXiv:2003.07000*.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2012.

[21] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech emotion recognition: Methods and cases study," in *Proc. 10th Int. Conf. Agents Artif. Intell.*, Madeira, Portugal, 2018, pp. 175–182, doi: 10.5220/0006611601750182.

[22] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Proc. Social Media Mach. Learn.*, London, U.K., 2019, pp. 1–4, doi: 10.5772/intechopen.84856.

[23] D. Jing, T. Manting, and Z. Li, "Transformer-like model with linear attention for speech emotion recognition," *J. Southeast Univ.*, vol. 37, no. 2, pp. 164–170, Jun. 2021, doi: 10.3969/j.issn.1003-7985.2021.02.005.

[24] D. Craft, *Dialectical Behavior Therapy: Control Your Emotions, Overcome Mood Swings and Balance Your Life With DBT*. Scotts Valley, CA, USA: CreateSpace Independent Publishing Platform, 2018.

[25] L. Huang, J. Dong, D. Zhou, and Q. Zhang, "Speech emotion recognition based on three-chanel feature fusion of CNN and BiLSTM," in *Proc. ICIAI*, Xiamen, China, 2020, pp. 52–58, doi: 10.1145/3390557.3394317.

[26] S. Ottl, S. Amiriparian, M. Gerczuk, V. Karas, and B. Schuller, "Group-level speech emotion recognition utilizing deep spectrum features," in *Proc. ICMI*, Amsterdam, The Netherlands, 2020, pp. 821–826, doi: 10.1145/3382507.3417964.

[27] D. Rana and A. Jain, "Effect of windowing on the calculation of MFCC statistical parameter for different gender in Hindi speech," *Int. J. Comput. Appl.*, vol. 98, no. 8, pp. 6–10, 2014.

[28] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Commun.*, vol. 52, nos. 7–8, pp. 613–625, 2010, doi: 10.1016/j.specom.2010.02.010.

[29] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7194–7198, doi: 10.1109/ICASSP40776.2020.9053581.

[30] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Appl. Acoust.*, vol. 179, Aug. 2021, Art. no. 108046, doi: 10.1016/j.apacoust.2021.108046.

[31] R. Rumagit, G. Alexander, and I. Saputra, "Model comparison in speech emotion recognition for Indonesian language," *Proc. Comput. Sci.*, vol. 179, pp. 789–797, Dec. 2021.

[32] O. U. Kumala and A. Zahra, "Indonesian speech emotion recognition using cross-corpus method with the combination of MFCC and teager energy features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 163–168, 2021.

[33] S. Shahsavarani, "Speech emotion recognition using convolution neural networks," M.S. thesis, Dept. Comp. Sci. Eng., Nebraska Univ., Lincoln, Nebraska, 2018.

[34] J. Rintala, "Speech emotion recognition from raw audio using deep learning," M.S. thesis, Dept. Comp. Sci. Eng., KTH Royal Inst. Tech., Stockholm, Sweden, 2020.

[35] D. M. Waqar, T. S. Gunawan, M. A. Morshidi, and M. Kartiwi, "Design of a speech anger recognition system on Arduino nano 33 BLE sense," in *Proc. IEEE 7th Int. Conf. Smart Instrum., Meas. Appl. (ICSIMA)*, May 2021, pp. 64–69, doi: 10.1109/ICSIMA50015.2021.9526323.

FELICIA ANDAYANI received the bachelor's degree in information and communication technology from the Swinburne University of Technology Sarawak Campus, Malaysia, in 2019, and the Master of Science (by Research) degree in artificial intelligence and speech emotion recognition research.

LAU BEE THENG (Senior Member, IEEE) has actively contributed to her research areas with various edited books, peer-reviewed journals, conference proceedings, higher degrees in research, and funded research projects. She is currently a Senior Member of the Association of Computing Machinery, a Professional Technologist, and a Certified Software Tester. Her research interests include artificial intelligence in activity recognition, natural scene text recognition, speech emotion detection, road accidents recognition, wafer surface defect detection, financial risks recognition, and aesthetic preference of design objects.

MARK TEEKIT TSUN (Member, IEEE) received the B.Sc. degree (Hons.) in computer science from Coventry University, in 2005, the master's degree in software engineering from OUM, and the B.Eng. (Hons.) and Ph.D. degrees from the Swinburne University of Technology Sarawak Campus, in 2014 and 2018, respectively. He worked in the software development industry, until 2008. He joined the Faculty of Engineering, Computing, and Science as a Lecturer, in 2019. His research interests include computer game development, drone technology applications, assistive robotics for injury prevention, human–robot interaction, assistive technologies, virtual, augmented, mixed reality, drone development, and the Internet of Things (IoT).

CASLON CHUA received the B.Sc., M.Sc., and Ph.D. degrees in computer science from De La Salle University, Manila, Philippines, in 1988, 1993, and 1999, respectively. He is currently the Acting Department Chair of Computing Technologies with the School of Software and Electrical Engineering, Swinburne University of Technology, Hawthorn, VIC, Australia. His research interests include computing education, data visualization, database systems, human–computer interactions, and software engineering.