A feature selection based parallelized CNN-BiGRU network for speech emotion recognition in Odia language

Bubai Maji
Silicon Institute of Technology, Bhubaneswar
Monorama Swain (mswain@silicon.ac.in)
Silicon Institute of Technology, Bhubaneswar
Rutuparna Panda
VSSUT

Research Article

Keywords: Speech emotion recognition, Self-attention, Fusion method, Deep-learning

Posted Date: April 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1529387/v1

Abstract

Emotion recognition from speech is an integral part of human interaction. This paper represents work that includes the creation and evaluation of speech emotion recognition on the Odia database. Previous research in this field implemented several benchmark datasets. In this work, we use two benchmark datasets for cross-validation with our own created Odia dataset name as SITB-OSED. Initially, the spectral, prosodic, and voice quality features are extracted from a raw audio file; secondly, a Gradient Boost Decision Tree (GBDT) feature selection method is used to remove all the redundant features and select the potential features. Here, two distinct series of experiments are performed. Firstly, the baseline model, which takes all the combined selected features, is chosen as input (spectral, prosodic, and voice quality features). Secondly, the proposed model processes all the selected features through two separate channels, the Convolutional neural network (CNN) and Bi-directional gated recurrent units (Bi-GRU). Specifically, the proposed method achieved 6.67%, 6.03%, and 5.55% higher accuracy than the baseline model on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Interactive Emotional Dyadic Motion Capture (IEMOCAP) and SITB-OSED datasets. We also report that the proposed parallelized CNN-BiGRU model outperforms the recent state-of-the-art methods with an accuracy of 82.29% and 78.54% on the RAVDESS and IEMOCAP datasets, respectively. For our SITB-OSED dataset, the overall recognition accuracy of 84.02% is achieved.

1 Introduction

Speech is a highly convenient and natural means for humans to communicate their emotions, perceptions, and intentions. For the past few decades, the Speech Emotion Recognition (SER) system has attracted much attention to speech processing research (Khalil et al. 2019). Its applications that improve human-computer interaction includes human-robot communication, the caller's voice emotional state requirements from a call center in case of emergency, level of satisfaction of a customer's identity, education, and medical analysis (Ramakrishnan et al. 2013; Zhang et al. 2020; Wani et al. 2021). Emotion recognition from speech signals is a complex task due to the fact that speech signals depend on various factors such as speaker, gender, dialect, and others (Alsharhan et al. 2020). Several efforts have been made to identify the emotional state of the speech signals.

To detect the speaker's emotional state, features are needed to be extracted from the speech signal. Therefore, many researchers have focused on searching for novel speech features that indicate different emotions. There are various distinguishing handcrafted features usually used for recognizing speech emotion. The qualitative features are formant frequency, harmonics-to-noise ratio (HNR), and energy. The prosodic features are pitch, fundamental frequency, duration, and some main spectral features are Log-frequency Cepstral Coefficients (LFPC), Linear Prediction Coefficients (LPC), Mel-frequency Cepstral Coefficients (MFCC) (Abdel-Hamid. 2020; Shivaprasad et al. 2021). It is still uncertain which features are more prominent, potential, and informative about emotions. Many hands-designed features have been used for speech emotion recognition (Zhang et al. 2021). However, some of them are low-level features and may not carry enough information to detect subjective emotions.

Deep neural networks (DNNs) providing a possible solution to overcome this issue include convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Abdul Qayyum et al. 2019; Zhu et al. 2020). Li et al. (2019) implemented a potential method using a 1D CNN to capture the emotional feature through extracting complementary features, achieving superior and competitive results. The evaluation was carried out by using three emotion corpus, which are IEMOCAP (Busso et al. 2008), RAVDESS (Livingstone and Russo 2018) and EMODB (Burkhardt et al. 2005). For the past few years, a lot of work has been successfully done on CNN and learning the feature from speech signals. The recurrent neural network model is used to pass the information sequentially with a time series step; in real life, the speaker's utterance is in the time-series form (Zhu et al. 2020). In recent years, many improvements to the internal structure of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been made. However, the RNN network is used to enhance memory information, but it has computational complexity. The LSTM and GRU networks become the most suitable choice for using temporal information of speech and successfully deploys for SER (Swain et al. 2021).

Research work in the field of SER task motivates us to use different deep learning models to learn more about deep features. Some issues need to be addressed to fulfil the goal of the SER tasks. First, speech signals may have a different time duration, and most models accept a fixed size of the input length. The models with the fixed size of input length may not be able to learn fewer emotional details to train a robust network. So, the ability of the individual neural network to recognize emotion is limited (Li et al. 2019; Wijayasingha et al. 2021).

To address these challenges, we present two different deep learning models using CNN and Bi-GRU network to improve speech emotion recognition. One model combines the CNN-BiGRU in a serial fashion and the other is a Parallel combination is proposed for SER task. The advantage of the proposed parallel CNN-BiGRU architecture is that it is easier to design and suitable for utterance-level features. The serial model structure may lose some emotional information due to the inheritance relationship between CNN and Bi-GRU during the training period.

Still, this work faces two challenges: i) Which model is faster and has less computationally complex; and ii) To identify the most valuable spectral, voice quality, and prosodic features for the SER task. The proposed parallel model consists of two channels; one is CNN and another, the Bi-GRU channel. The CNN part of the model extracts local features of the selected features to preserve each data's emotional detail. The model has a stronger ability to feature learning. At the same time, the selected utterance-level features are pass through a Bi-GRU network to learn the contextual semantics information of speech signals with different duration of speech samples and overcome the loss of temporal information. Then fused all high-level features learned in Bi-GRU and CNN models. Finally, fused features are classified by the Softmax classifier. To train and evaluate the two models, first, we use our own Odia dataset. Subsequently, two benchmark datasets (RAVDESS and IEMOCAP) are used on the same models for cross-validation of the Odia dataset.

In this part, it is summarize the contribution of the suggested work. The next module consists of a brief discussion of the previous work related to speech emotion recognition. The dataset and the methodology of multiclass speech emotion recognition using the baseline and proposed model are described in section 3. The results of our experiments follow in section 4. Finally, the conclusion and future work conclude in section 5.

2 Literature Review

Deep learning has created an enormous impact on building a reliable SER system, even with a small dataset considering relevant features. As there is a challenge in creating large datasets with added computational costs, the researchers try to improve the performance of the classification using different deep learning methods. Most of them use the typical deep learning networks like convolutional neural networks, recurrent networks, and combinations (Passricha et al. 2020; Issa et al. 2020; Swain et al. 2021). However, some researchers have also used the Gaussian mixture model (GMM) and Hidden Markov model (HMM) using the robust features to recognize the emotional state from speech (Anagnostopoulos et al. 2015). First, Tzirakis et al. (2017) implemented an end-to-end deep learning model for the SER system by combining the CNN and multilayer LSTM to extract contextual information from raw audio signals. The authors achieved good results compared to the previous models of that time. A dual-level LSTM model was proposed to combine handcrafted and natural audio signal features (Wang et al. 2020). The authors used different time-frequency resolutions to pre-process each utterance into handcrafted input and two Mel-spectrograms on the IEMOCAP database.

Many researchers have also used bi-directional LSTM and GRU models to learn more long-time contextual information in two directions as a basic structure (Yu and Kim 2020). It has been recently demonstrated that attention-based deep learning model can significantly improve classification results. The attention mechanism is applied to both the LSTM and GRU network outputs. It is reasonable because the improved gate is only related to the previous cell state and is independent of the present input, which could reduce the computational complexity (Zhu et al. 2020). Some experimental results reveal that a new attention-gate-based model can also improve the recognition rate. For example, Li et al. (2021) proposed Bi-LSTM with self-attention to analyze the correlation between speech signal and emotion to increase the gap of emotional information.

Recently, Chen and Huang (2021) used a dual attention-based Bi-LSTM mechanism is used with hybrid features for SER and tested on the IEMOCAP dataset. Various approaches have been discussed in the literature for the recognition of speech emotion. Some researchers were also using the multi-hop attention and fusion methods have been adopted for further developing the SER systems. The multi-hop attention mechanism is used to capture the relevant information on textual data, which is consequently applied to attend each segment for classification. A multi-hop attention model has been developed to combine acoustic data and textual information for the SER task on the IEMOCAP dataset (Yoon et al. 2019).

Wang (2020) utilized a multi-feature fusion method to improve the local attention of contextual and textual information. Sometimes, combined prosodic and spectral features give better recognition in an SER system. Also, the head fusion and self-attention-based multi-head mechanism can improve the performance of the SER system (Xu et al. 2021). Experiments were conducted with attention-based convolutional neural networks on the IEMOCAP dataset, with four (neutral, angry, sad, and excited) emotions using MFCCs features. Another confidence-based fusion mechanism was developed using three multi-task models: DNN, CNN, and RNN using utterance-level HSF (high statistical function), multiple segment-level Mel-spectrograms, and multiple frame-level LLDs (low-level descriptors). They are used for four (angry, happy, neutral, and sad) categorical discrete emotion recognition (Yao et al. 2020). As speech is a time-series signal, it could be modeled as a temporal sequence of the local acoustic representation to obtain SER performances. Shirian and Guha. (2021) employ the graph-based convolutional model and trained on IEMOCAP and MSP-IMPROV datasets to improve the utterance level SER.

Furthermore, another important goal in the speech emotion recognition field is identification of important features for training the model. A new one-dimensional deep CNN framework is built with the combinations of MFCCs, Chromagram, Mel-scaled spectrogram, Spectral contrast, and Tonnetz representation features using RAVDESS and IEMOCAP datasets (Issa et al. 2020). Recently, Hou et al. (2022) proposed a multi-view model to integrate shared and view-specific features to determine the relationship between heterogeneity and consistency information for SER. The author used three types of LLDs and pass them into different modules, then fused the features obtained by the three modules and achieved competitive results on IEMOCAP and Emo-dB datasets.

3 Methods Of Multiclass Speech Emotion Recognition

The performance of the SER system is mainly dependent on the methodology of modeling and the relations between the emotional labels and speech features across various languages. Many different audio datasets are available in public like EMO-DB (535 utterances), IEMOCAP (10039 utterances), RAVDESS (1440 sample), EMOVO (588 utterances), etc. (Burkhardt et al. 2005; Busso et al. 2008; Livingstone and Russo 2018; Costantini et al. 2014), are widely used for the research of SER systems. However, emotional analysis of the Odia language is not often encountered. In this research work, an audio dataset on Odia language is created apart from using two other benchmark datasets, RAVDESS, and IEMOCAP used for cross-validation of our Odia dataset. The Odia language is an Indo-Aryan language spoken in various parts of eastern India. This language has been influenced by the neighboring family of languages, both Dravidian and Aryan. The dialectal variations are Baleswari (Balasore), Sambalpuri (Sambalpur and western districts), Cuttacki (Cuttack), Beharampuri (Beharampur), Ganjami (Ganjam and Koraput area), Chhattisgarhi (Chhattisgarh and Odisha adjoining areas), and Medinipuri (Midnapur district of West Bengal and Odisha border area) (Jain and Cardona 2007). After presenting the datasets, the process of feature extraction, feature selection followed by the baseline proposed model, is described.

3.1 Datasets

The SITB-OSED database contains 7317 utterances recorded from 12 Odia speakers (6 male and 6 female) in Odia Indic language. They express six different states of emotions- happiness, surprise, anger, fear, sadness, and disgust. Every emotion includes almost equal utterances to measure the classification more accurately. The waveform of seven emotions is shown in Fig. 1. All the utterances were recorded with a sampling rate of 22050 Hz and 16-bit quantization. For the extraction of features, utterances are compressed by 16000 Hz. Table 1 shows the number of audio samples in each emotion in our SITB-OSED dataset. The dataset will be publicly available at: https://www.speal.org/sitb-osed/.

Table 1
The number of emotion samples in the SITB-OSED dataset

Emotion	Number	Participation (%)
Anger	1197	16.36
Disgust	1231	16.82
Fear	1196	16.34
Happiness	1197	16.36
Sadness	1309	17.89
Surprise	1187	16.22

RAVDESS is an English language dataset. This dataset presents the recording of 24 actors (12 female and 12 male) with 1440 utterances. The utterances consist of eight (8) different emotions (angry, fear, disgust, sadness, calm, happiness, neutral, and surprise).

Here the IEMOCAP dataset is used as a second benchmark database for the experiment. This dataset is categorized into two parts, scripted dialogs, and improvised dialogs. The whole dataset contains five sessions recorded by ten actors with a total duration of 12 hrs. Each session includes recordings of two (2) actors (1 female and 1 male) with a sampling rate of 16000 Hz. The total audio samples of the dataset are classified by nine different emotions (anger, excitement, happiness, frustration, sadness, disgust, neutral, surprise, and fear). In this experiment, we use only four emotion classes (anger, neutral, happiness, and sadness) of improvised samples. Most of the previous works using IEMOCAP utilizes these four emotion labels.

3.2 Feature extraction

The feature extraction process is vital for building a successful deep neural network; for this study, 70 features are extracted, including voice quality, prosodic, and spectral features, for analysis. Spectral features were extracted using the Python Librosa audio library (McFee et al. 2015), and prosodic and voice quality features were extracted using Praat software, contain 13 MFCCS, 16 LPC, and 41 prosodic

and voice quality features. The prosodic and voice quality features include prominent features like HNR, Fundamental frequency (mean and standard deviation), Formant frequency (Minimum and Maximum of first formant frequency to fifth formant frequency (F1 – F5)), Formant bandwidth (Mean, Standard deviation and Median of first formant bandwidth to fifth formant bandwidth (B1 – B5)), Jitter (local, rap, local absolute, ppp5, ddp), Shimmer (local, localdb, apq3, apq5, apq11, dda) and PCA (shimmer and jitter). apq3, apq5, apq11, dda) and PCA (shimmer and jitter).

3.3 Feature selection

The performance of any model or system depends mainly on the number of inputs or input features. Here a feature selection method incorporated before the classification task to get relief from correlated features. Because a set of large number features includes irrelevant features that give the model over-fitting and reduce the accuracy. Selecting a small number of features allows for a better model fitting and helps the performance.

In this work a GBDT feature selection mechanism used to obtain the best feature for emotion classification (Wijaya et al. 2019; Zulfiqar et al. 2021). It is a statistical approach and provides an essential score of each feature, calculating how much linear dependency exists between a pair of features. Gradient boosting has low variance and high bias. This method has a computational complexity, but gives higher accuracy than other feature selection methods such as recursive elimination, forward and backward methods. In this algorithm, 75% data is used for training.

The input dataset for training is defined as $\{u_i, v_i\}^P$ (where p is the total number of sample), $\left(u_i\right)$ is the input vector in the form of $\{u_1, u_2, \ldots, u_i\}$, and $\left(v_i\right)$ is the corresponding class of emotions $\{v_1, v_2, \ldots, v_i\}$. This algorithm is used to ensure the convergence of the GBDT. The basic learner is $h(u)$, where $u_i = (u_{1i}, u_{2i}, \ldots, u_{mi})$. m is the number of predicted variables. The normalized multiclass GBDT for the p^{th} sample is given by,

$$\Psi_n = \underset{\Psi}{\operatorname{argmin}} \sum_{i=1}^{p} L(v_i, F_{n-1}\left(u_i\right) + \Psi h_n(u_i))$$

1

Where, $L(v, F(u))$ is the differentiable loss function and n is the times of iteration.

The pseudo code of GBDT is described as follows:

1. Input data: $\{u_i, v_i\}^P$ with total number of iterations(k)

2. Initialize result with a selected feature set. The initial constant value of the model Ψ is given:

$$F_0(u) = \underset{\Psi}{\operatorname{argmin}} \sum_{i=1}^{p} L(v_i, \Psi), i = \{1,2, \ldots . , p\}$$

2

3. Iteration starts from 1 to k

i) Update the weights for targets based on previous output.

ii) Fit the model on selected sample of dataset.

4. Update the weights for targets based on previous output.

$$F_n(u) = F_{n-1}(u) + \Psi_n h_n(u)$$

3

5. Return the final output with final iterations of $F_k(u)$.

From the above feature selection based algorithm the 30 best features which contribute more than all other features for recognition accuracy. Table 2 shows the details of the 30 best-selected features, including the spectral, voice quality, and prosodic features which is used in the models.

Table 2
Selected features after feature selection method

Feature	Group
MFCCs: 0 to 10 and 12; LPCs: 1,2,3,5,7,8,11;	Spectral features
Mean (f_0), Standard deviation (f_0), HNR, Local Jitter, Max (F_3), Min (f_5), Median (B_1, B_2), apq11, Shimmer, Mean (B_5), dda Shimmer;	Combined prosodic and voice quality features

3.4 **Proposed baseline model**

In the proposed baseline model, a cascading CNN-Bi GRU model with a self-attention layer for emotion classification from the speech feature of a raw audio file is used. The baseline model consists of two parts, the first part is a one-dimensional convolutional layer that cascades with the second part that is the Bi-GRU layer, followed by a self-attention layer and fully connected network. The suggested proposed baseline model is shown below (Fig. 2).

3.4.1 Convolutional neural network

First, the experiments are conducted using pre-trained AlexNet networks (Krizhevsky et al. 2012) instead of CNN. The pre-trained AlexNet model required large data because it trained on millions of image data. In this experiment, the pre-trained model did not get good accuracy because of a limited number of speech data. After that, our own CNN network is created. We ran several simulations of models randomly with various convolutional layers, kernel sizes, kernel filters, and other hyper-parameters, and selected the optimal number of convolutional layers and other parameters those values give more accurate results.

The above-proposed framework consisted of four 1D convolutional layers. We choose a 1D convolutional network over a 2D because the 1D network allows larger feature datasets. Moreover, the learning rate of sequential data is faster than the 2D CNN model (Kiranyaz et al. 2019). Each convolutional layer is followed by batch normalization, pooling layer, and a dropout layer. The dataset size was S×R (where S represents the number of samples and R is denoted the final dimension of the feature).

The first convolutional layer received 30×1 as the input of the model. A filter size of 256 is used in the initial convolutional layer with a kernel size of 5×1. After that, batch normalization and activation layer are used. The batch normalization layer allows one to learn independently of each layer and is activated by linear rectifier units ('Relu'). Then a max-pooling layer is applied to compress the data and reduce the complexity; to map the input data, the size of the pooling layer was 2. After the pooling operation, one dropout layer is used to minimize the over-fitting problem, which helps better the model's recognition rate. The dropout value set at 0.4, which means 40% of the neurons are dumped during the training process. This same process is repeated four times with different filter sizes and kernel sizes. The output of the last convolutional layer is connected to the bidirectional gated recurrent unit.

3.4.2 Bi-directional gated recurrent unit

The Bi-directional gated recurrent unit is an enhancement type of recurrent neural network (Zhu et al. 2020). It has two RNNs paths; one is forward, and the other is a backward path and concatenated with the same output layer; for an audio sound which is a frequency-based time-varying speech signal, the Bi-GRU layer to add two independent hidden layers, one forward, and another in the backward direction is used.

For i^{th} utterances of the input feature vector u with encoded by time step t is mathematically described in two layers as follows:

$$\overrightarrow{h_{i,t}} = \overrightarrow{GRU}\left\{\left(u_{i,t}, \overrightarrow{h_{i,t-1}}\right) + \overrightarrow{b_f}\right\}$$

4

$$\overleftarrow{h_{i,t}} = \overleftarrow{GRU}\left\{\left(u_{i,t}, \overleftarrow{h_{i,t+1}}\right) + \overleftarrow{b_b}\right\}$$

5

$$\hat{y}_i = \overrightarrow{h}_{i,t} + \overleftarrow{h}_{i,t}$$

6

Where, $\overrightarrow{h}_{i,t}$, $\overleftarrow{h}_{i,t}$ and $\hat{y}_i$ is the forward path, backward path, and output state. Similarly $u_{i,t}$, $\overrightarrow{h}_{(i,t-1)}$, $\overleftarrow{h}_{(i,t+1)}$ is the input vector of i^{th} utterance at time t, forward hidden state at a time step of $(t-1)$, and backward hidden state at a time step of $(t+1)$ corresponding with the forward bias ($\overrightarrow{b_f}$) and backward bias ($\overleftarrow{b_b}$). This baseline framework uses two Bi-GRU layers because they experimentally perform better than one Bi-GRU layer.

The first Bi-GRU layer is connected to the input of the last convolutional layer with 64 neural units. The second Bi-GRU layer is serially connected to the output of the first Bi-GRU layer with 32 neurons, and the output was activated of each layer by the same activation function 'Relu' followed by the dropout rate of 0.4. The baseline combines CNN-BiGRU model loses some long-term-dependencies features when features pass through CNN layers. To overcome the problems and improved the performance significantly two separated CNN and Bi-GRU channels are used in the suggested proposed model.

3.4.3 Self-attention layer

After the passing of important information of feature through the output of the Bi-GRU layer, a self-attention mechanism is constructed to capture further important information (Li et al. 2021). The self-attention layer cannot forget the past information, and it can respond more accurately to understand the sequence of information. Then, a flattened layer is used to get a fixed length of weight matrix passed through a dense layer with 64 hidden units. Finally, the final dense layer contains 8 neurons for the RAVDESS dataset, 7 neurons for the Odia dataset, and 4 neurons for the IEMOCAP dataset, equal to the number of emotions in the different datasets. Then the emotion is classified using the 'Softmax' classifier. The baseline model uses the 'Categorical Cross-Entropy' as a loss function and 'Adam' optimizer with a learning rate of 0.0001 and a decay rate of 1e-06.

3.5 Proposed model

In this paper, proposed baseline model describes the basis for the final proposed model. We modify the baseline model by changing, adding, and removing some layers to better recognize the emotion classification task. The characteristics discussed in the final proposed model are given in the next section. Figure 3 shows the final proposed model.

The first part of the proposed model consists of two parallel convolutional layers with a different kernel size of (15×1) and (5×1) to extract vertical (spectral, voice quality, and prosodic features) and the horizontal (cross-time) form. Then a concatenate layer is used to map the selected input feature. After

mapping the input feature, two parallel deep learning networks are used; one is a convolution network another is a Bi-GRU network. The convolution network is composed of three convolutional layers with the filter size of 128, 256, and 256 corresponds to the kernel size of (5×1), (7×1), and (7×1). These three convolutional layers pass all the features from the input feature maps. CNN learns the frequency domain spectral features (MFCCs, LPCs) more accurate than the time series feature. In each convolutional layer, a batch normalization layer is used with an activation function 'Relu'. After that a max-pooling layer is used to reduce layer complexity and maps the feature weight. Finally, the global average poling (GAP) layer is employed to reduce the number of training parameters and generates feature points for fusion. It can also overcome the over-fitting problem.

The second channel of the parallelized CNN-BiGRU model consists of two Bi-GRU layers with an equal number of neurons which is used in our baseline model. This separated Bi-GRU layer is used to easily learn the time-series features in a feature vector sequence (Zhu et al. 2020). So the prosodic and voice quality features pass quickly with the minimum loss of information and learn more high-level time-series features through the Bi-GRU layer from the input feature map. And, with the help of the self-attention layer, some long-term dependencies are also learned and all information is passed through the flatten layer to reshape the feature vector for fusion operation.

3.5.1 Fusion method

Typically, the fusion method uses fusion features as input. To combine various neural networks to obtain an efficient prediction result, a confidence-based decision-level method is used similar to decision score fusion to get the final classification results. A confidence-based decision-level fusion technique takes the outputs of every recognizer as associate input and confirms which combination of features gives a better result (Yao et al. 2020; Wang et al. 2021). The confidence score from the model is obtained from the convolutional module, and the Bi-GRU model is used as the one-dimensional confidence score vector. These confidence score vectors are summed up together. The mathematical expression of the fusion vector is as follows:

$$w_1 = \{c_1, c_2, c_3, \ldots . , c_n\}$$

7

$$w_2 = \{d_1, d_2, d_3, \ldots\ldots, d_n\}$$

8

$$w^{fusion} = f^{sum}(w_1, w_2)$$

9

Here, w_1 and w_2 are the output of the spectral feature vector of the convolutional network, the output of the prosodic feature vector of the Bi-GRU. The w^{fusion} represents the fusion vector. After the feature is

fused, the fused vector passes through the dense layer for the final prediction. The number of neurons in the dense layer and other hyper-parameters were the same as in the previous model. And, proposed model shows that the recognition rate is better than the baseline model.

4 Experimental Results And Discussion

This research work was implemented using the TensorFlow backend and Keras deep learning platforms (Geron 2017). In this study, SITB-OSED dataset and benchmark datasets both were split randomly using the k-fold cross-validation. Here a five-fold ($k = 5$) cross-validation technique is used. All the data from each dataset are split with a ratio of 0.8:0.1:0.1 (train/validation/test) according to the speakers. So 80% of the data used for training the model, 10% for validation, and the rest 10% is used to test the model performance for the recognition rate of emotion. A similar experimental result shows the proposed model, which gives a greater recognition rate than the baseline model. The Precision, Recall, F-score, and overall accuracy are used to evaluate the overall recognition rate of each emotion category on SITB-OSED dataset and other two benchmark databases, in the form of a confusion matrix.

4.1 Experimental results of the baseline model

Tables 3 through 5 demonstrate the statistical performance of each emotion in terms of Precision, Recall, and F-score of the baseline model. The result indicates that different datasets present different difficulties in identifying a particular emotion. For a detailed analysis of the recognition rate of the baseline model, the confusion matrix is demonstrated as shown in Figs. 4 through 6.

Table 3
The performance (%) measure of the baseline model for each emotion of the SITB-OSED dataset

Emotion	Precision	Recall	F-score
Anger	72.15	88.19	79.36
Disgust	77.08	74.79	75.91
Fear	80.24	75.80	77.95
Happiness	74.51	78.54	76.46
Sadness	93.22	86.72	89.84
Surprise	78.61	66.79	72.21

Table 4
The performance (%) measure of the baseline model for each emotion of the IEMOCAP dataset

Emotion	Precision	Recall	F-score
Anger	62.38	60.42	61.38
Happiness	61.63	64.91	63.22
Neutral	77.05	79.39	78.20
Sadness	80.34	85.37	82.77

Table 5
The performance (%) measure of the baseline model for each emotion of the RAVDESS dataset

Emotion	Precision	Recall	F-score
Anger	93.11	69.23	79.41
Calm	79.44	89.47	84.15
Disgust	74.53	64.86	69.35
Fear	80.16	73.68	76.78
Happiness	68.05	74.29	71.03
Neutral	75.28	75.00	75.13
Sadness	78.61	78.95	78.77
Surprise	69.84	79.49	74.35

Figure 4 shows the performance of the baseline model on the SITB-OSED dataset. It is found that anger emotion and sadness are classified well with an accuracy of 88.19% and 86.72% from this confusion matrix. In contrast, surprise is hard one to recognize as compare to other emotions with an accuracy of 66.79%, and the other three emotions are classified relatively well with than 80%. Also it is being observed that the emotions of disgust are mostly confused with high recognition neutral emotion.

Figure 5 demonstrates that sadness and neutral emotion perform well with an accuracy of 85.37% and 79.39%. At the same time, happiness and anger perform relatively low with an accuracy of 64.91% and 60.42%. And also, observed that anger misclassified 35.42% with neutral emotion and 21.05% happiness misclassified with anger emotions on the IEMOCAP dataset.

The confusion matrix on the RAVDESS dataset is shown in Fig. 6. From Fig. 6, it is observed that the baseline model identified calm with a high recognition rate, and fear, happiness, neutral, sadness, and

surprise were identified moderately well. Then the recognition rate of anger and disgust are low, whereas fear is confused with sadness emotion.

4.2 Experimental results of proposed best model

Tables 6 through 8 show the recognition rate of each emotion in the form of precision, recall, and F-score of the second proposed model.

Table 6
The performance (%) measure of the proposed model for each emotion of the SITB-OSED dataset

Emotion	Precision	Recall	F-score
Anger	80.41	85.43	82.84
Disgust	82.18	86.36	84.21
Fear	81.72	80.82	81.26
Happiness	83.56	84.47	84.01
Sadness	89.77	92.97	91.32
Surprise	84.39	74.09	78.89

Table 7
The performance (%) measure of the proposed model for each emotion of the IEMOCAP dataset

Emotion	Precision	Recall	F-score
Anger	67.29	68.75	68.01
Happiness	81.57	73.68	77.42
Neutral	79.08	80.70	79.88
Sadness	78.44	91.06	84.28

Table 8
The performance (%) measure of the proposed model for each emotion of the RAVDESS dataset

Emotion	Precision	Recall	F-score
Anger	88.09	71.79	79.10
Calm	82.17	92.11	86.85
Disgust	84.39	72.97	78.26
Fear	85.64	78.95	82.15
Happiness	79.41	86.84	82.95
Neutral	91.72	91.67	91.69
Sadness	76.37	76.32	76.34
Surprise	77.04	87.18	81.80

Here, we discuss the performance of each emotion of the proposed model on our Odia dataset and two benchmark datasets, which show relatively better performance than the baseline model and previous framework results. We demonstrate the confusion matrix of the proposed model as shown in Figs. 7 through 9.

Figure 7 shows the performance level of each emotion of the proposed model on the SITB-OSED dataset. The confusion matrix indicated the predicted label corresponding to the actual label. From the confusion matrix of the SITB-OSED dataset, we see that the recognition rate of disgust from 74.79–86.36%, fear from 75.80–80.82%, happiness from 78.54–84.47%, sadness is increased from 86.72–92.97%, and surprise from 66.79–74.09% then the baseline model. Only anger emotion identifies same as baseline model. This report shows the advantages of the proposed model over the baseline model.

Figure 8 demonstrates the confusion matrix of our model on the IEMOCAP dataset. The results indicate that anger, happiness, and sadness, emotion are significantly improved by 8.33%, 8.77%, 4.88%, and 7.59% respectively. And sadness performs almost the same as baseline model.

The confusion matrix on the RAVDESS dataset of the second proposed model is illustrated in Fig. 9. The experimental results indicate that the proposed model leads to disgust from 64.86–72.97%, fear from 73.68–78.95%, happiness from 74.29–86.84%, neutral emotion from 75–91.67%, and surprise from 79.49–87.18%, and recognition rate of other three emotions are relatively same as baseline approach with 82.29% average recall rate (un-weighted accuracy).

So, the overall report shows that the proposed model performs superiorly over the baseline model and recent some state-of-art works.

4.3 Comparison with state-of-the-art models

Here a novel method is adopted for the SER task to improve the recognition rate compared to the previous methods. Table 9 shows that the proposed model performs significantly better as compared to state-of-the-art models on the RAVDESS dataset. For instance, experimental result presents the improved levels of 12.89%, 10.68%, 5.27%, 6.53%, and 4.42% compared with (Jalal et al. 2019; Issa et al. 2020; Mustaqeem et al. 2020; Fu et al. 2021; Muppidi el al. 2021) respectively. Jalal et al. (2019) and Mustaqeem et al. (2020) adopted different models such as a hybrid model using Bi-LSTM and 1D convolutional-capsule network, and a deep Bi-LSTM network for SER system using spectrogram. Issa et al. (2020) extracted spectral features to input their deep convolutional model to detect speech emotions.

On the IEMOCAP dataset, the proposed method is also compared with four recent approaches, as illustrated in Table 10. Our model clearly outperforms with an increasing rate of 8.14%, 6.29%, 8.08%, and 4.34% (Zhang et al. 2019; Mustaqeem et al. 2020; Muppidi el al. 2021; Gat et al. 2022) with the overall accuracy level of 78.54%.

To this end, Table. 11 presents the performance of the proposed model corresponding to the conventional deep learning methods and also the baseline method on the SITB-OSED dataset. As there was no such work has done previously on the Odia dataset, here two benchmark datasets used for cross-validation of our created Odia dataset. As a result, SITB-OSED dataset shows an overall recognition rate of 84.02%, which exhibits better performance than conventional methods as well as baseline method as shown in Table 11.

Table 9
Performance (%) comparisons of the proposed method with state-of-the-art models in the RAVDESS dataset

Model (Refs.)	**Accuracy (%)**
Jalal et al. (2019)	69.40
Issa et al. (2020)	71.61
Mustaqeem et al. (2020)	77.02
Xu et al. (2021)	77.80
Muppidi el al. (2021)	77.87
Our baseline model	75.62
Our Parallelized CNN-BiGRU model	**82.29**

Table 10
Performance (%) comparisons of the proposed method with state-of-the-art models on the IEMOCAP dataset

Model (Refs.)	Accuracy (%)
Zhang et al. (2019)	70.40
Mustaqeem et al. (2020)	72.25
Muppidi el al. (2021)	70.46
Gat et al. (2022)	74.20
Our baseline model	72.51
Our Parallelized CNN-BiGRU model	**78.54**

Table 11
Performance (%) comparisons of the method with conventional method on the SITB-OSED dataset

Model (Refs.)	Accuracy (%)
CNN + LSTM	74.85
CNN + GRU	73.91
Our baseline model	78.47
Our Parallelized CNN-BiGRU model	**84.02**

5 Conclusions And Future Direction

This research work presents one baseline proposed model and one parallelized CNN-BiGRU model to develop a better system for emotion recognition from speech utterances. First, the methodology of how our proposed model learns the best spectral, voice quality, and prosodic features information after the feature selection process, using GBDT from an audio signal has been discussed. Then, in the proposed model, CNN is used to learn the frequency-based spectral feature for producing high label segment features, and two Bi-GRU layers are deployed to the model to learn long-term dependency features. Finally, the output of the two-channel features is fused for the final classification through a fully connected network.

The performance of the two models was trained and tested on SITB-OSED dataset and two benchmark databases (RAVDESS and IEMOCAP). Here, it is shown that the proposed parallelized CNN-BiGRU model has a distinct advantage over the proposed baseline model in terms of overall recognition rate. Also, this model achieves better recognition speed than the serial CNN-BiGRU model. The parallel model learns the input features simultaneously, and the serial CNN-BiGRU model learns the input features one after another. So, the parallelized model required significantly less time than the serial model. The experimental

results reveal that the proposed method outperforms the state-of-the-art methodology on the RAVDESS and for the IEMOCAP datasets. Furthermore, on the SITB-OSED dataset, we compares the proposed method to a set of conventional deep learning models and achieved superb performance. The proposed method can also overcome the shortcomings of limited speech emotion datasets, which is a roadmap for developing a deep learning model for SER. The model trained over 150 epochs which show the stability of the model. Therefore, an end-to-end machine learning methodology can improve performance.

For continued research, in the future we aim to investigate the performance of the proposed model with different Odia dialects. Furthermore, an end-to-end based deep learning methodology and also multi-label feature extraction will be explored to further delineate the emotion recognition task.

Declarations

Author Contributions: Bubai Maji designed, analysis, interpretation of the data and wrote the manuscript, Monorama Swain, Rutuparna Panda revising it critically for important intellectual content, and approval of the final version, Monorama Swain supervised the project.

Funding: This work is supported by the DST, Govt. of India, under grant reference no. DST/ICPS/CLUSTER/Data Science/2018/General, Date: 07/01/2019.

Declaration of Competing Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment: This research work has been supported by the Department of Science and Technology (DST) under the Grant no. DST/ICPS/CLUSTER/Data Science/2018/General. The authors are indebted for the input given by Prof. J. Talukdar in improving the manuscript.

References

1. Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2), 155–177
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *In: Ninth European conference on speech communication and technology.*
3. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S. … Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335–359
4. Costantini, G., Ladarola, I., Paoloni, A., & Todisco, M. (2014). EMOVO Corpus: An Italian emotional speech database. *In: Proceedings of the 9th international conference on language resources and evaluation*, pp 3501–3504
5. Geron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. California: O'Reilly Media Inc

6. Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894
7. Jain, D., & Cardona, G. (2007). *The Indo-Aryan languages*. London: Routledge
8. Jalal, M. A., Loweimi, E., Moore, R. K., & Hain, T. (2019). Learning temporal clusters using capsule routing for speech emotion recognition. *In: Interspeech*, pp 1701–1705
9. Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2019). 1D convolutional neural networks and applications: a survey.arXiv PreprintarXiv:1905.03554.
10. Krizhevsky, A., & Sutskever, I. (2012). Hinton, & G. E. Imagenet classification with deep convolutional neural networks. *In: Advances in neural information processing systems*, pp 1097–1105. https://dl.acm.org/doi/pdf/10.1145/3065386
11. Li, D., Liu, J., Yang, Z., Sun, L., & Wang, Z. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert System with Applications*, 173, 114683. https://doi.org/10.1016/j.eswa.2021.114683
12. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): https://doi.org/10.5281/zenodo.1188976
13. Li, Y., Baidoo, C., Cai, T., & Kusi, G. A. (2019). Speech emotion recognition using 1D CNN with no attention. *In: 2019 23rd international computer science and engineering conference (ICSEC)*, pp 351–356
14. McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: audio and music signal analysis in python. *In: Proceedings of the 14th python in science conference*, pp 18–25
15. Mustaqeem., Sajjad, M., & Kwon, S. (2020). Clustering based speech emotion recognition by incorporating learned features and deep Bi-LSTM. *IEEE Access*, 8, 79861–79875. https://doi.org/10.1109/ACCESS.2020.2990405
16. Passricha, V., & Aggarwal, R. K. (2020). A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR. *Ambient Intelligence and Humanized Computing*, 11, 675–691. https://doi.org/10.1007/s12652-019-01325-y
17. Ramakrishnan, S., & Emary, I. M. M. (2013). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 52(3), 1467–1478
18. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8), 1301–1309
19. Wang, C. (2020). Speech emotion recognition based on multi-feature and multi-lingual fusion.arXiv preprintarXiv:2001.05908.
20. Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., & Tarokh, V. (2020). Speech emotion recognition with dual-sequence LSTM architecture. *In: 2020 IEEE international conference on acoustic, speech and signal processing (ICASSP)*, pp 6474–6478

21. Wang, S. H., Nayak, D. R., Guttery, D. S., Zhang, X., & Zhang, Y. D. (2021). COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis. *Information Fusion*, 68, 131–148
22. Wijaya, A., Kharis, & Prastuti, W. (2019). Gradient boosted tree based feature selection and parkinson's disease classification. *In: 2019 5th international conference on science and technology*
23. Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Communication*, 120, 11–19
24. Yoon, S., Byun, S., Dey, S., & Jung, K. (2019). Speech emotion recognition using multi-hop attention mechanism. *In: 2019 IEEE international conference on acoustic, speech and signal processing (ICASSP)*, pp 2822–2826
25. Yu, Y., & Kim, Y. J. (2020). Attention-LSTM-attention model for speech emotion recognition and analysis of iemocap database. *Electronics*, 9(5), 713–725
26. Zhang, J., Zhou, Y., & Liu, Y. (2020). EEG-based emotion recognition using an improved radial basis function neural network. *Ambient Intelligence and Humanized Computing*, 20, 1–12
27. Zhang, S., Tao, X., Chuang, Y., & Zhao, X. (2021). Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Communication*, 127, 73–81
28. Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y. (2019). Attention based fully convolutional network for speech emotion recognition.arXiv preprintarXiv:1806.01506.
29. Zhu, Z., Dai, W., Hu, Y., & Li, J. (2020). Speech emotion recognition model based on Bi-GRU and focal loss. *Pattern Recognition Letters*, 140, 358–365
30. Muppidi, A., & Radfar, M. (2021). Speech emotion recognition using quaternion convolutional neural networks. *In: ICASSP 2021*, pp. 6309–6313
31. Gat, I., Aronowitz, H., Zhu, W., Morais, E., & Hoory, R. (2022). Speaker normalization for self-supervised speech emotion recognition.arXiv preprintarXiv:2202.01252v1.
32. Zulfiqar, H., Yuan, S. S., Huang, Q. L., Sun, Z. J., Dao, F. Y., Yu, X. L., & Lin, H. (2021). Identification of cyclin protein using gradient boost decision tree algorithm. *Computational and Structural Biotechnology Journal*, 19, 4123–4131
33. Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*, 7, 117327–117345
34. Swain, M., Maji, B., & Das, U. (2021). Convolutional Gated Recurrent Units (CGRU) for Emotion Recognition in Odia Language. *IEEE EUROCON 2021–19th International Conference on Smart Technologies, 2021*, pp. 269–273
35. Alsharhan, E., & Ramsay, A. (2020). Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition. *Language Resources and Evaluation*, 54, 975–998
36. Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, 9, 47795–47814

37. Shivaprasad, S., & Sadanandam, M. (2021). Dialect recognition from Telugu speech utterances using spectral and prosodic features. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-021-09854-8
38. Abdel-Hamid, L. (2020). Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features. *Speech Communication*, 122, 19–30
39. Abdul Qayyum, A. B., Arefeen, A., & Shahnaz, C. (2019). Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. *In Proceedings of the IEEE International Conference on Signal Processing, Information, Communication Systems, 2019*, pp. 122–125
40. Wijayasingha, L., & Stankovic, J. A. (2021). Robustness to noise for speech emotion classification using CNNs and attention mechanisms. *Smart Health*, 19, 100165
41. Chen, Q., & Huang, G. (2021). A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. *Engineering Applications of Artificial Intelligence*, 102, 104277
42. Xu, M., Zhang, F., & Zhang, W. (2021). Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset. *IEEE Access*, 9, 74539–74549
43. Shirian, A., & Guha, T. (2021). Compact Graph Architecture for Speech Emotion Recognition. *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021*, pp. 6284–6288
44. Hou, M., Zhang, Z., Cao, Q., Zhang, D., & Lu, G. (2022). Multi-View Speech Emotion Recognition via Collective Relation Construction. *In IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 218–229

Figures

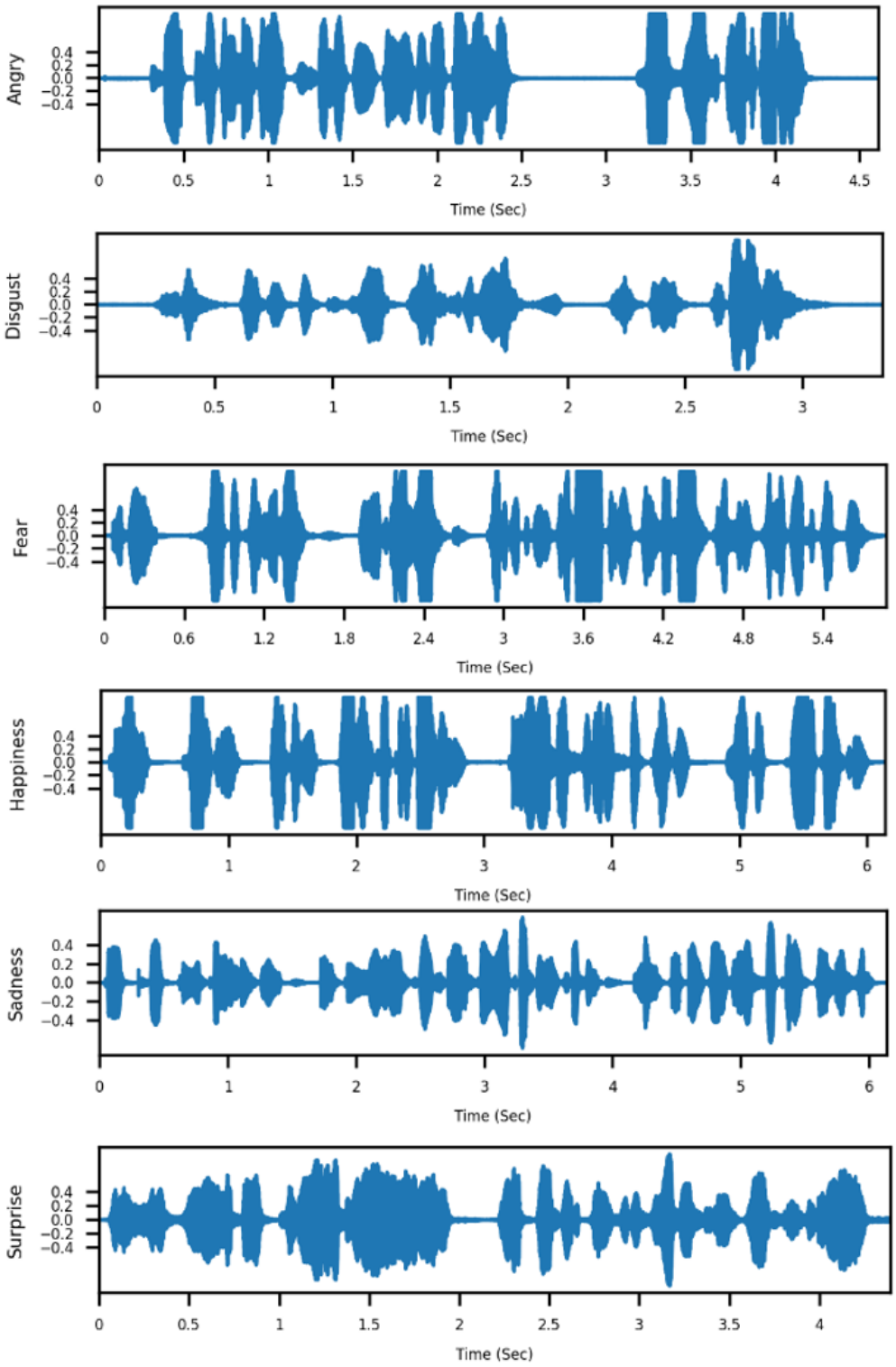

Figure 1

The waveform of each emotion on SITB-OSED dataset

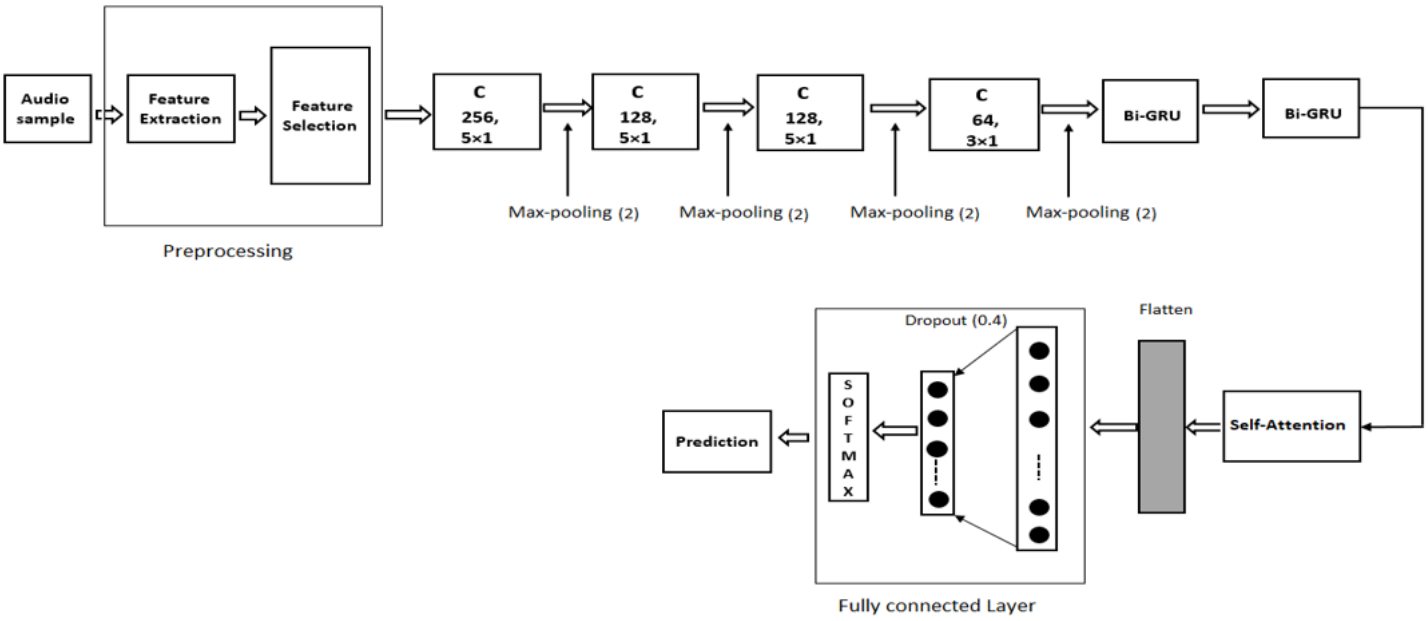

Figure 2

The baseline proposed architecture

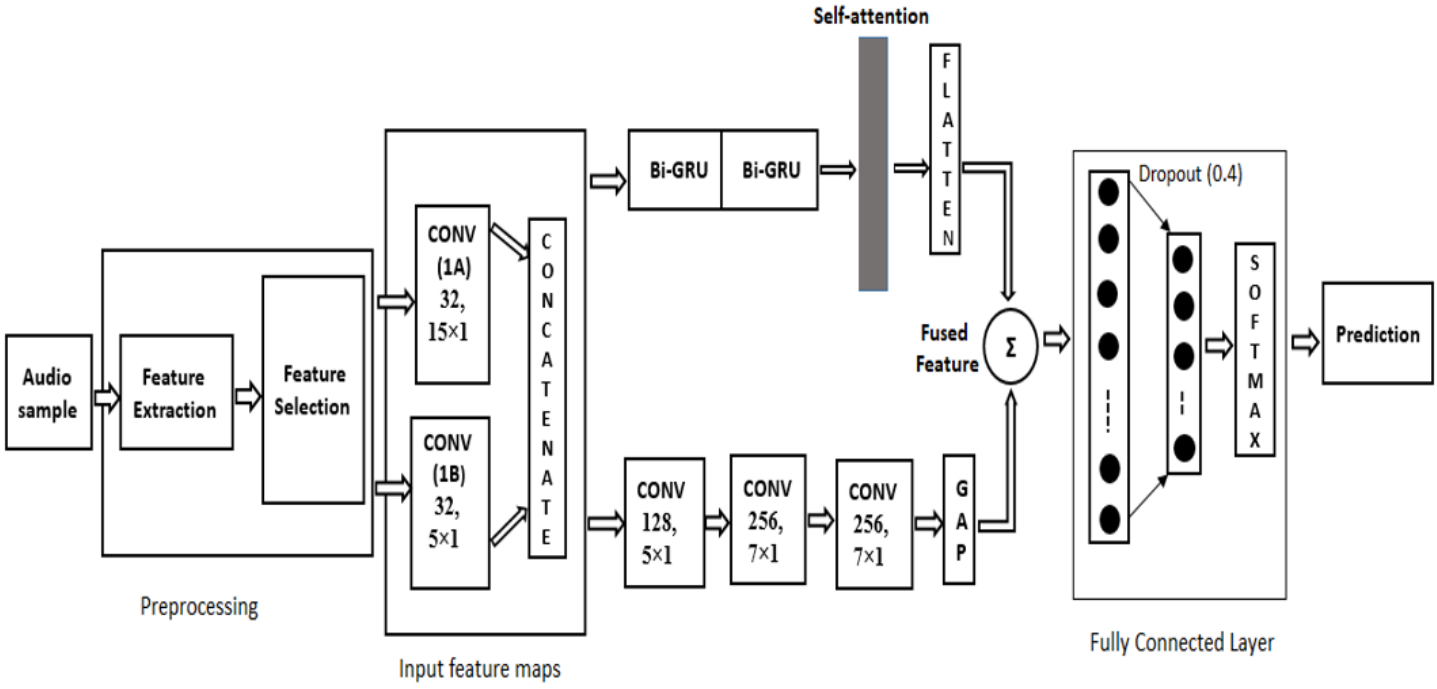

Figure 3

Proposed parallelized CNN-BiGRU model

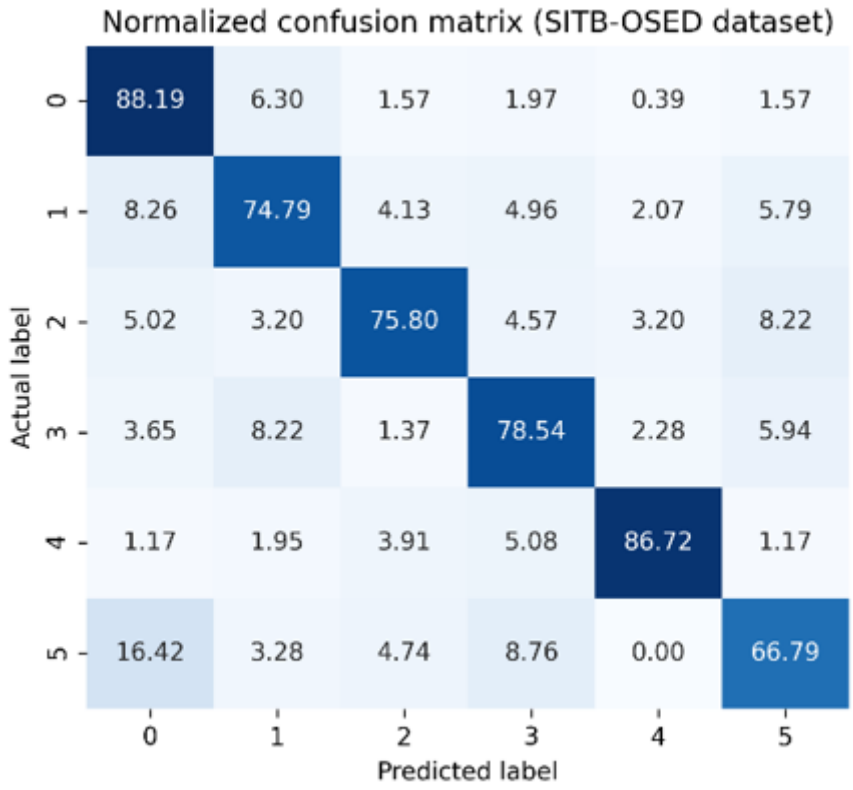

Figure 4

Confusion matrix of prediction result of baseline model with an accuracy of 78.47% on the SITB-OSED dataset (0-anger; 1- disgust; 2-fear; 3-happiness; 4-sadness; 5-surprise)

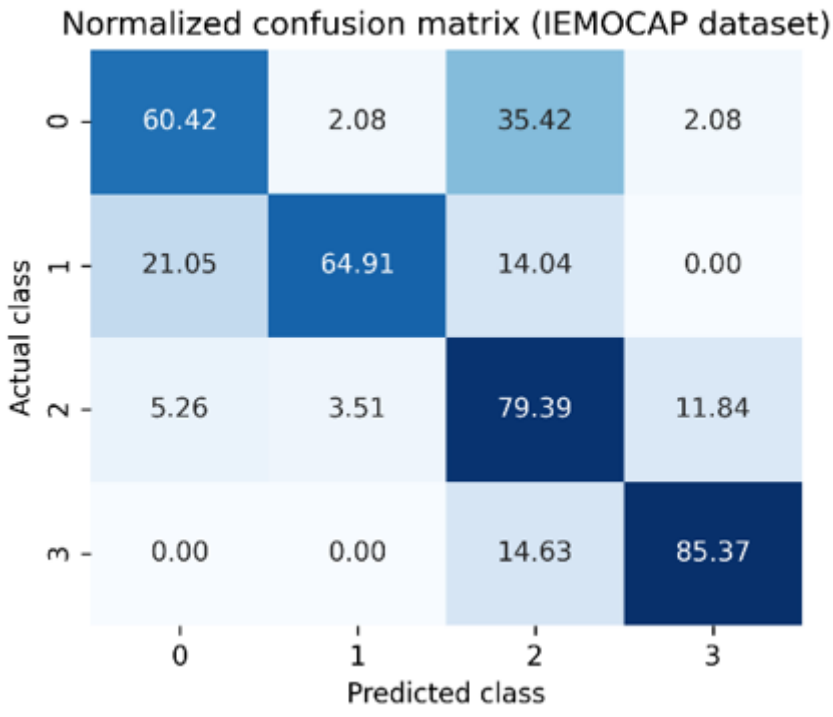

Figure 5

Confusion matrix of prediction result of baseline model with an accuracy of 72.51% on the IEMOCAP dataset (0-anger; 1-happiness; 2-neutral; 3-sadness)

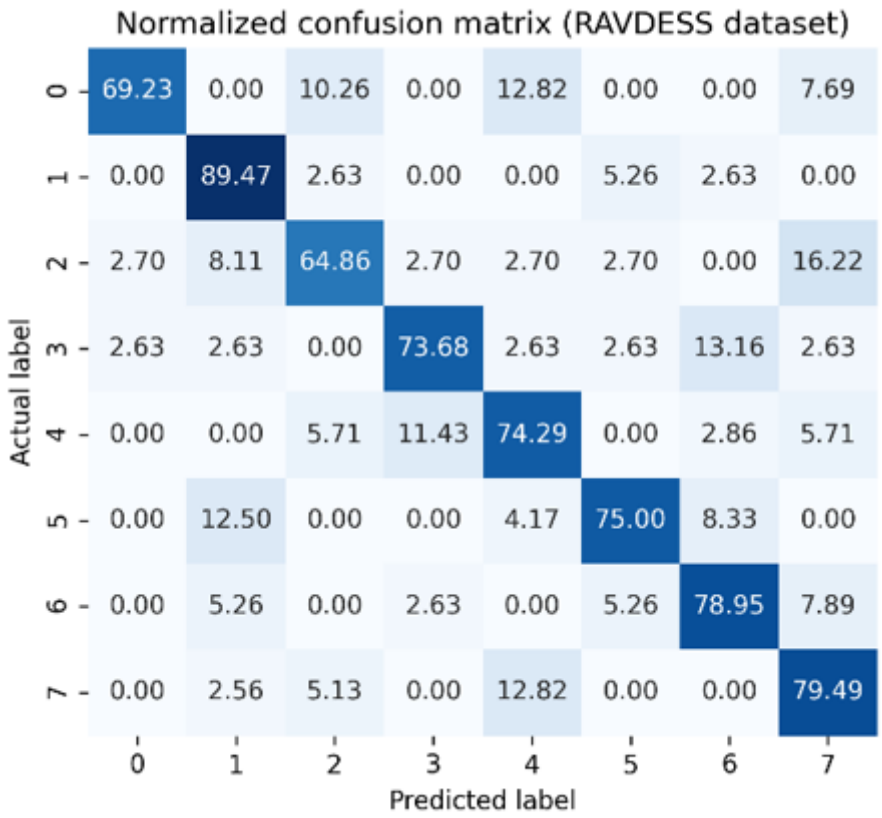

Figure 6

Confusion matrix prediction result of baseline model with an accuracy of 75.62% of the RAVDESS dataset (0-anger; 1-calm; 2-disgust; 3-fear; 4-happiness; 5-neutral; 6-sadness; 7-surprise)

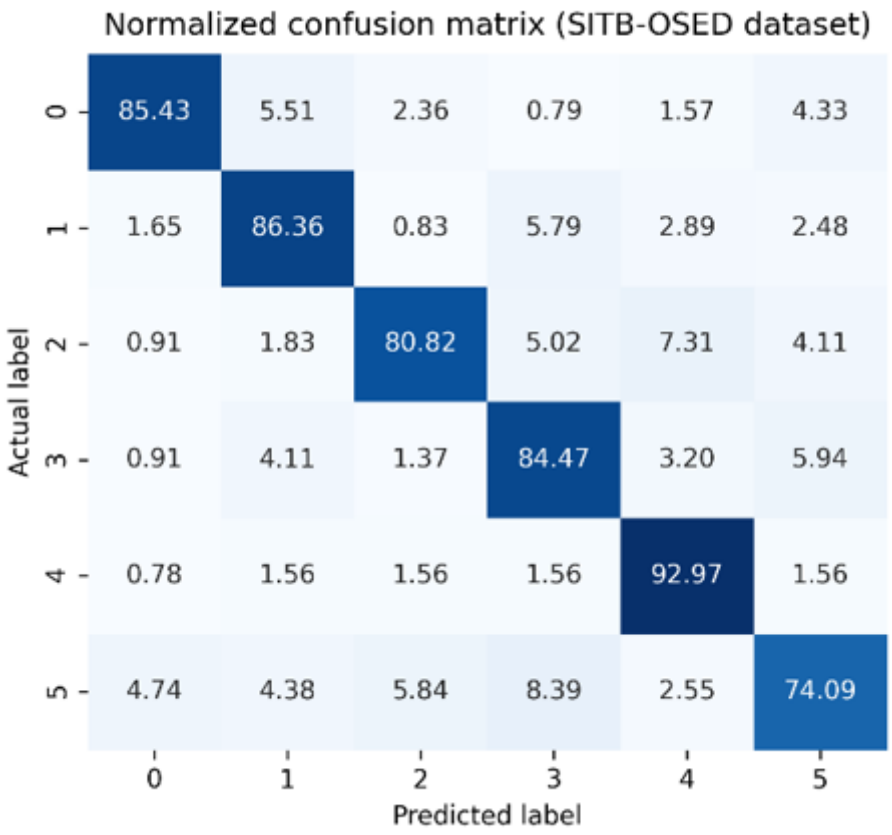

Figure 7

Confusion matrix prediction result of proposed model with an accuracy of 84.02% on the SITB-OSED dataset (0-anger; 1-disgust; 2-fear; 3-happiness; 4-sadness; 5-surprise)

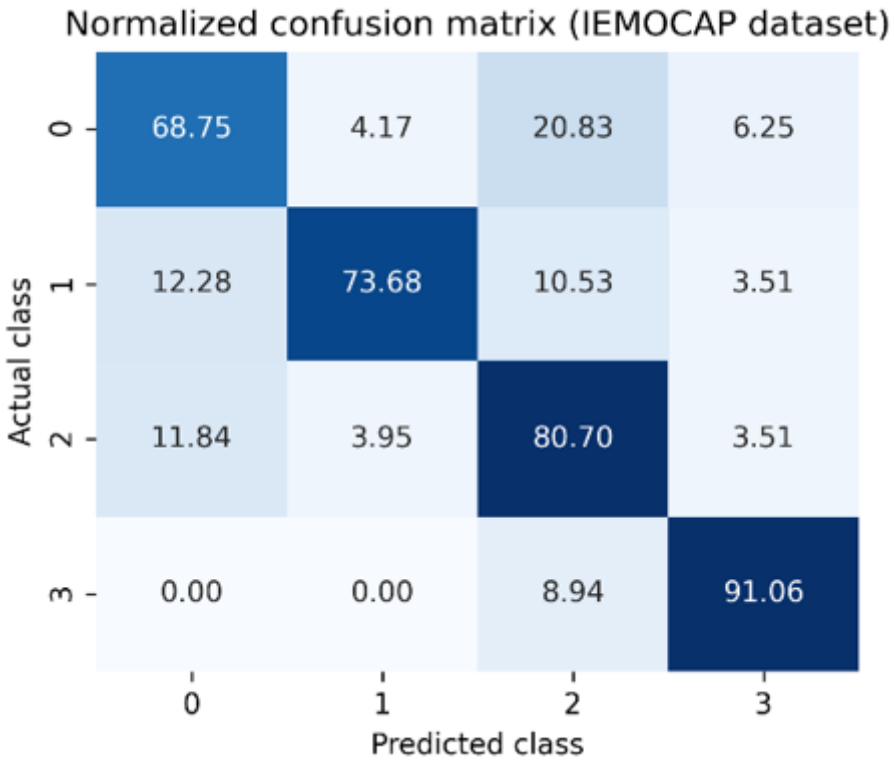

Figure 8

Confusion matrix prediction result of proposed model with an accuracy of 78.54% on the IEMOCAP dataset (0-anger; 1-happiness; 2-neutral; 3-sadness)

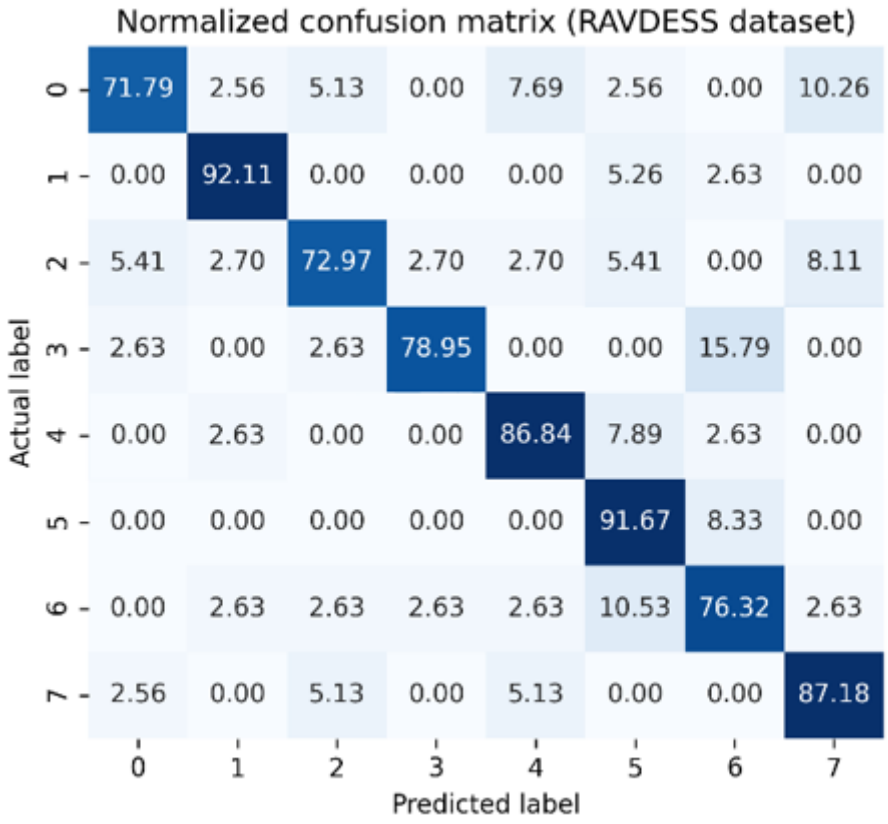

Figure 9

Confusion matrix prediction result of proposed model with an accuracy of 82.29% on the RAVDESS dataset (0-anger; 1-calm; 2-disgust; 3-fear; 4-happiness; 5-neutral; 6-sadness; 7-surprise)