Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer

Report

Science and Technology Advancement- Assessment Review (STAAR)

Rana Sujeet Kumar
Regno.2101020306

under the supervision of
Dr. Ashish Ranjan

Department of Computer Science and Engineering
C.V. Raman Global University, Bhubaneswar
Odisha, 752054, India

Abstract

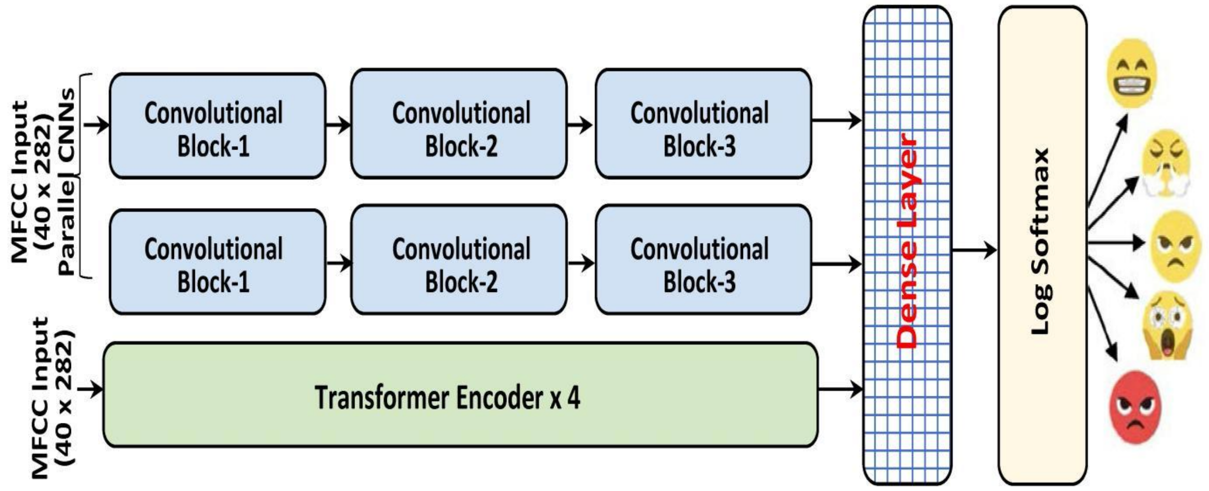
Speech emotion recognition (SER) is a challenging task in human–computer interaction
(HCI) systems. One of the key challenges in speech emotion recognition is to extract the emotional features effectively from a speech utterance. Despite the promising results of recent studies, they generally do not leverage advanced fusion algorithms for the generation of effective representations of emotional features in speech utterances. To address this problem, we describe the fusion of spatial and temporal feature representations of speech emotion by parallelizing convolutional neural networks (CNNs) and a Transformer encoder for SER. We stack two parallel CNNs for spatial feature representation in parallel to a Transformer encoder for temporal feature representation, thereby simultaneously expanding the filter depth and reducing the feature map with an expressive hierarchical feature representation at a lower computational cost. We use the RAVDESS dataset to recognize eight different speech emotions. We augment and intensify the variations in the dataset to minimize model overfitting. Additive White Gaussian Noise (AWGN) is used to augment the RAVDESS dataset. With the spatial and sequential feature representations of CNNs and the Transformer, the SER model achieves 82.31% accuracy for eight emotions on a hold-out dataset. In addition, the SER system is evaluated with the IEMOCAP dataset and achieves 79.42% recognition accuracy for five emotions. Experimental results on the RAVDESS and IEMOCAP datasets show the success of the presented SER system and demonstrate an absolute performance improvement over the state-of-the-art (SOTA) models.

1 Introduction

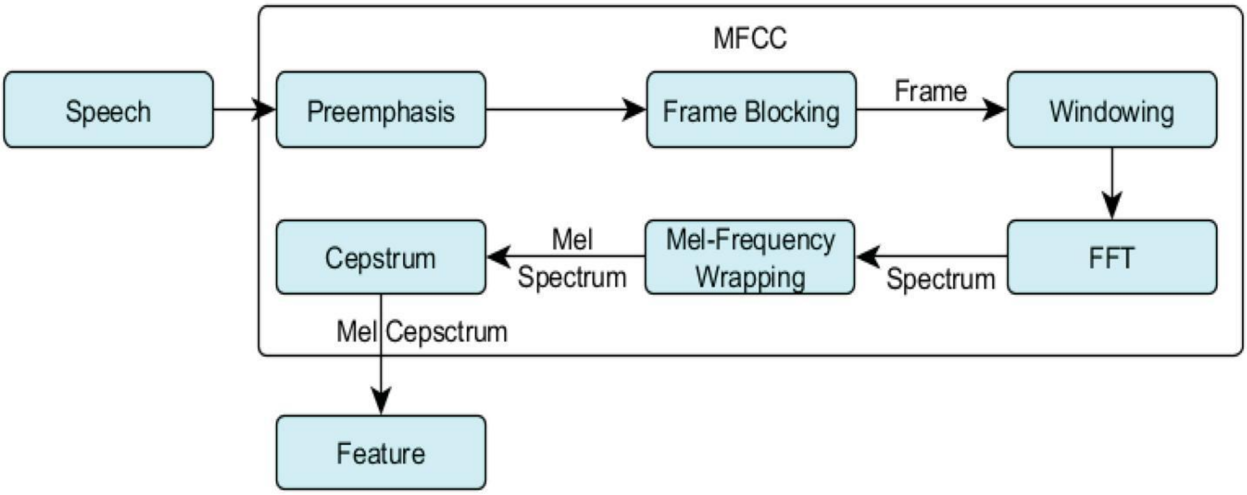
Speech Emotion Recognition (SER) is a crucial task in human-computer interaction systems, enabling machines to interpret and respond to human emotions more effectively. A key challenge in SER is the extraction of emotional features from speech utterances. While recent studies have shown promising results, many of them do not utilize advanced fusion algorithms for generating effective emotional feature representations. In this research, we propose a fusion of spatial and temporal feature representations of speech emotion by parallelizing Convolutional Neural Networks (CNNs) and a Transformer encoder for SER. Our proposed model, named CTENet, consists of two parallel CNNs for spatial feature representation and a Transformer encoder for temporal feature representation. By stacking two parallel CNNs, we expand the filter depth and reduce the feature map, resulting in an expressive hierarchical feature representation at a lower computational cost. The Transformer encoder is utilized to learn the context-dependent (temporal) features, allowing the model to capture emotional patterns that vary with the situational context. We evaluate the performance of our proposed model using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. Data augmentation techniques, such as adding white noise, are used to increase the variability in the datasets and minimize model overfitting. Our experimental results demonstrate that the CTENet model achieves state-of-the-art performance on both datasets, outperforming existing models for SER. The proposed model has potential applications in various fields, including online tutorials, language translation, intelligent driving, and therapy sessions, where the ability to recognize and respond to human emotions is essential.

2 Related work

Speech Emotion Recognition (SER) is a challenging research area that involves recognizing human emotions from speech signals. The SER system consists of two main modules: feature representation and emotion classification. The feature representation module extracts relevant features from the speech signal, while the emotion classification module identifies the emotional state based on these features. In recent years, various techniques have been proposed to improve the performance of SER systems. One popular approach for feature representation is Mel-Frequency Cepstral Coefficients (MFCCs), which are used in [49] to classify various emotions using the logistic model tree (LMT) classifier. Another approach is to use pre-trained Convolutional Neural Network (CNN) architectures, such as AlexNet and VGG, to extract features from spectrograms using transfer learning [52]. A trained CNN model can also be used for feature extraction, followed by Support Vector Machine (SVM) for emotion classification [53]. In addition to MFCCs, other features such as prosodic and spectral features can also be used for SER. For instance, [54] uses 1D-CNN + FCN-based SER to classify various speech emotions using prosodic and spectral features from



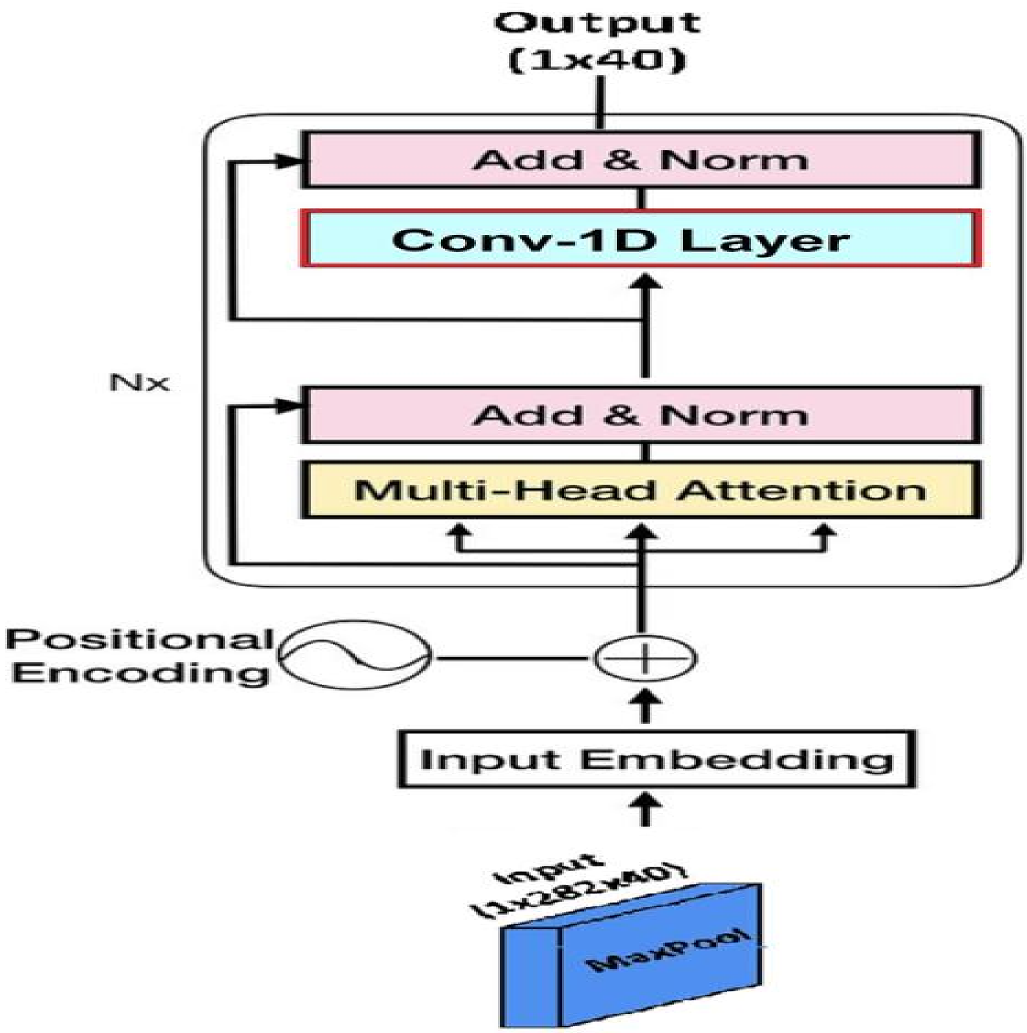
MFCCs. Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) can also be used to classify long-term sequences in speech signals for SER [55]. Recently, deep learning models have gained popularity in SER due to their ability to learn complex features from raw speech data. The DNN-LSTM-based SER method [56] uses a hybrid approach to learn spatiotemporal cues from raw speech data. The CNN-BLSTM-based SER method [57] learns the spatial features and temporal cues of speech symbols and increases the accuracy of the existing model. The SER extracts spatial features and feeds them to the BLSTM to learn temporal cues for the recognition of the emotional state. Attention mechanism-based deep learning for SER is another notable approach that has achieved vast success. In classical DL-based SER, all features in a given utterance receive the same attention. However, emotions are not consistently distributed over all localities in the speech samples. In attention-based DL, attention is paid by the classifier to the given specific localities of the samples using attention weights assigned to a particular locality of data. The SER system based on multilayer perceptron (MLP) and a dilated CNN [68] uses channel and spatial attention to extract cues from input tensors. The proposed CTENet SER system consists of two parallel CNNs and a multi-head attention Transformer encoder to recognize emotions in speech spectrograms. The CNN modules with skip connections extract spatial features, while the Transformer encoder captures temporal cues. The extracted features are fed to a fully connected dense network for emotion classification. The CTENet model achieves state-of-the-art performance on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets.



3 Proposed Model

This section experimentally examines the proposed CTENet model for SER and demonstrates its efficiency. We conducted extensive experiments by using the standard REVDESS dataset, an acted speech emotions dataset for SER. In addition, the IEMOCAP dataset was used to examine the performance across different databases. The performance of the proposed CTENet model has been evaluated with other state-of-the-art (SOTA) SER models that are reported in the recent literature. We also conducted an ablation learning study to confirm the multi-head attention performance in the CTENet model for SER. A complete description of the speech emotion datasets, model training/testing/validation, and emotion recognition output with discussion is given in the following sections.

The CTENet model for SER provides outscored results in terms of emotion recognition using MFCC spectrograms. The proposed CTENet model was tested over two benchmark speech emotion datasets (RAVDESS and IEMOCAP). The speech signals were transformed into MFCC coefficients representing an utterance as a grayscale image, an appropriate 2D representation for CNN models. Adam was used to optimize the model, with a cross-entropy loss function for 200 epochs. Utterance-level extensive experiments were performed to observe the significance of the CTENet model. We followed a 80%–20%–20% training/testing/validation ratio during the experiments. Various evaluation metrics were used to examine the prediction performance of the models, such as accuracy, the F1 score, precision, and recall. We trained the CTENet models on two datasets and examined them from different aspects to demonstrate their advantages.



For the comparison, we selected the following SOTA baseline models to extensively evaluate the performance of the CTENet model. Att-Net [68] is a robust SOTA lightweight self-attention model for SER, where a dilated CNN uses channel and spatial attention for the extraction of cues from the input tensors. The SVM ensemble model with a Gaussian kernel [50] is a standard benchmark used for SER comparison. The 1D-CNN [74] architecture is used, which extracts MFCC features and uses the trained 1D-CNN for emotion identification. The context-aware representations are used for emotion recognition. DeepNet [60] learns deep features and employs a plain rectangular filter with a new pooling scheme to achieve more effective emotion discrimination. The other SOTA models include GResNets [85]; SER using 1D-Dilated CNN, which is based on the multi-learning trick (MLT) [86]; and the CNN-BLSTM-based SER method from [57].

The proposed CTENet model demonstrated improved generalization during the experiments and evaluations for both datasets, and it obtained better emotion recognition accuracy with a low computational cost. In brief, we can assume that the proposed CTENet model for SER is accurate and computationally less complex. Consequently, it is able to examine human behaviors and emotions. Moreover, with the lightweight framework, this model is appropriate for real-time applications since it requires less training time. Table 9 gives the training time and model size (in Mb). We compared the training time and model size with those of other SER frameworks, including DS-CNN [51], CB-SER [57], and AttNet [68], for comparison. The experiments proved that the CTENet model is lightweight

(compact model size of 4.54 Mb), generalizable, and computationally less expensive, and it requires less processing time to recognize emotions, which indicates the appropriateness of the model for real-world applications. The processing time is significantly minimized as the simultaneously expanded filter depth and feature map reduction provide an expressive hierarchical feature representation at the minimum computational cost. The total trainable parameters are 222,248 for the CTENet model.

4 Performance Evaluation and Discussion

Table 10. Comparison of CTENet with benchmarks.

			RAVDESS Dataset			IEMOCAP Dataset		
Ref#	Benchmarks	Input Features	Accuracy	Precision	F1 Score	Accuracy	Precision	F1 Score
[56]	BE-SVM	Spectral Features	75.69	74.00	73.34	-	-	-
[85]	GResNets	Spectral Features	64.48	65.32	63.11	-	-	-
[86]	MLT-DNet	Spatial Features	-	-	-	73.01	74.00	73.00
[57]	Deep-BLSTM	Spatial + Temporal	77.02	76.00	77.00	72.50	73.00	72.00
[74]	1D-CNN	Spectral Features	71.61	-	-	64.30	-	-
[66]	DS-CNN	Spatial Features	79.50	81.00	84.00	78.75	86.00	82.00
[60]	DeepNet	Spatial + Temporal	-	-	-	77.00	76.00	76.00
[68]	Att-Net	Spatial Features	80.00	81.00	80.00	78.00	78.00	78.00
Our	CTENet	Spatial + Temporal	82.31	81.75	84.37	79.42	74.80	82.20

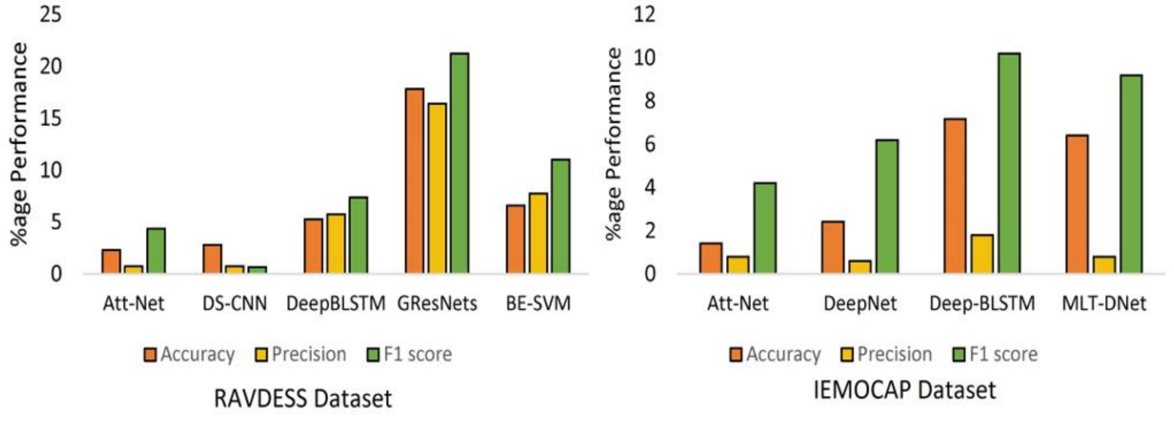


Figure 10. CTENet performance over SOTA for RAVDESS and IEMOCAP datasets.

The CTENet model is shown in Table 10, with improved accuracy and F1 scores over most SOTA benchmarks on both the RAVDESS and IEMOCAP datasets. CTENet outperforms models like GResNets, DeepNet, and 1D-CNN on the RAVDESS dataset. On the IEMOCAP dataset, CTENet achieves better performance than models like DeepNet, MLT-DNet, and Att-Net, with a 7.0% improvement in accuracy, 7.0% in precision, and 6.0% in F1 score.

5 Conclusions

we describe the combination of spatial and temporal feature representations of speech emotions by parallelizing CNNs and a Transformer encoder for SER. We extract the spatial and temporal features with parallel CNNs and the Transformer encoder from the MFCC spectrum. In the CTENet model, MFCCs are used as grayscale images, where the width is the time scale and height is the frequency scale. The experimental results on two popular benchmark datasets, RAVDESS and IEMOCAP, validate the usefulness of the CTENet model for SER. Our model achieves better experimental results over state-of-the-art models for speech emotion recognition, with overall accuracy of 82.31% and 79.80% for the benchmark datasets. Furthermore, the experimental results for different speech emotion classes show the effectiveness of the spatial and temporal feature fusion. The experimental results show the importance of MHAT inclusion in CTENet, where the emotion recognition results are improved significantly. The experimental results also prove that CTENet is compact (4.54 Mb) and computationally less costly, and requires less processing time to recognize different emotions, indicating the appropriateness of CTENet for real-world applications. With few entries in the datasets, the model sometimes overfits; however, we can fine-tune the model to avoid overfitting, such as by applying dropout regularization. It is also recommended to increase the database entries for better results and optimized model parameters.

The present study provides acceptable accuracy; however, a further improvement in accuracy can be achieved if the model architecture is further refined, e.g., a more effective feature extractor can be adopted. Different feature sets can be combined for more robust training features. Further, besides temporal and spatial features, we aim to add modalities to further increase the recognition accuracy using modality cues. In addition, we will apply recently introduced models to achieve state-of-the-art SER results

References

1. Liu, Z.T.; Xie, Q.; Wu, M.; Cao, W.H.; Mei, Y.; Mao, J.W. Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing 2018, 309, 145–156. [CrossRef]

2. Nwe, T.L.; Foo, S.W.; De Silva, L.C. Speech emotion recognition using hidden Markov models. Speech Commun. 2003, 41, 603–623. [CrossRef]

3. Patel, P.; Chaudhari, A.; Kale, R.; Pund, M. Emotion recognition from speech with gaussian mixture models via boosted gmm. Int. J. Res. Sci. Eng. 2017, 3, 294–297.

4. Chen, L.; Mao, X.; Xue, Y.; Cheng, L.L. Speech emotion recognition: Features and classification models. Digit. Signal Process. 2012, 22, 1154–1160. [CrossRef]

5. Koolagudi, S.G.; Rao, K.S. Emotion recognition from speech: A review. Int. J. Speech Technol. 2012, 15, 99–117. [CrossRef]

6. Akçay, M.B.; Ogˇuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 2020, 116, 56–76. [CrossRef]

7. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Survey of deep representation learning for speech emotion recognition. IEEE Trans. Affect. Comput. 2021, 14, 1634–1654. [CrossRef]

8. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Netw. 2017, 92, 60–68. [CrossRef]

9. Tuncer, T.; Dogan, S.; Acharya, U.R. Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowl.-Based Syst. 2021, 211, 106547. [CrossRef]

10. Singh, P.; Srivastava, R.; Rana, K.P.S.; Kumar, V. A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowl.-Based Syst. 2021, 229, 107316. [CrossRef]

11. Magdin, M.; Sulka, T.; Tomanová, J.; Vozár, M. Voice analysis using PRAAT software and classification of user emotional state. Int. J. Interact. Multimed. Artif. Intell. 2019, 5, 33–42. [CrossRef]

12. Huddar, M.G.; Sannakki, S.S.; Rajpurohit, V.S. Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN. Int. J. Interact. Multimed. Artif. Intell. 2021, 6, 112–121. [CrossRef]