International Journal of Speech Technology
https://doi.org/10.1007/s10772-023-10035-y

An efficient speech emotion recognition based on a dual-stream CNN-transformer fusion network

Mohammed Tellai[1] · Lijian Gao[1] · Qirong Mao[1,2]

Received: 12 December 2022 / Accepted: 29 May 2023

Abstract

The use of machine learning and artificial intelligence enables us to create intelligent systems. Speech emotion recognition system analyzes the speaker's speech to determine his/her emotional state. Speech emotion recognition is a challenging pattern recognition task. This paper proposes a new robust and lightweight speech emotion recognition system based on a dual-stream CNN-Transformer fusion network that effectively captures spatial and temporal information disentangled in long-distance raw features by exploring both MFCCs and Mel-spectrograms in a parallel mechanism. Experiments are performed on the widely used emotional benchmark datasets. Our approach has shown to be quite efficient. It outperforms the best-known state-of-the-art models with an accuracy of 97.64%, 99.42%, and 97.53% on RAVDESS, TESS, and EMO-DB datasets respectively. The results demonstrate the significant advantages of the proposed model and the ability of this architecture to learn to recognize emotional data features accurately.

Keywords Speech emotion recognition · Mel-spectrogram · MFCC · CNN · Transformer · Dual-stream model

1 Introduction

The goal of human-computer interaction (HCI) research is to provide a natural and effective interface for interactions between humans and computers. Furthermore, it aims to offer an appealing design, and a pleasurable user experience, support personal development, enhance online education, and so on. Emotions have naturally become a fundamental part of the development of HCI-based systems, as they are an essential part of social interactions (Gomathy, 2021; Liu, 2022).

Emotional content analysis of speech signals is a major research topic. Human emotions can be detected automatically through a variety of modalities, such as facial expressions, voice acoustics, hand gestures, and different movements of body parts. However, Speech is the most common means for humans to convey information, thus it is the primary option to detect emotion in the lack of visual information (facial expression and motion), such as in phone calls, audio messages, and contact center applications, therefore it is crucial to human–machine communication and interface engineering (Daneshfar et al., 2020). Emotions conveyed through signals should be reliably processed and appropriately handled to provide a more natural and intuitive interconnection between humans and machines. Over the past two decades, a variety of machine learning techniques have been proposed and improved with a focus on automatic emotion recognition.

Speech variations produced by participating speakers, in addition to diverse speaking styles, speaking velocity, and phrase structures make the speech emotion recognition (SER) task difficult and complex (El Ayadi et al., 2011; Singh & Goel, 2022). Another obstacle is that different people express different emotions differently based on context, culture, and surrounding circumstances. Different methods based on various speech features and deep learning

✉ Qirong Mao
mao_qr@ujs.edu.cn

Mohammed Tellai
5103180322@stmail.ujs.edu.cn

Lijian Gao
2112008002@stmail.ujs.edu.cn

[1] Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China

[2] Jiangsu Engineering Research Center of Big Data Ubiquitous Perception and Intelligent Agriculture Applications, Zhenjiang 212013, Jiangsu, China

Published online: 06 July 2023

techniques have been proposed by researchers. Pitch, formants, energy, linear frequency cepstral coefficient (LFCC), mel frequency cepstral coefficient (MFCC), and linear predictor coefficients (LPC), along with other acoustic features can be applied to the study of speech emotions. Numerous techniques have been employed to improve the performance of emotion classifiers. Machine learning methods, such as K-nearest neighbors (Morrison et al., 2007), Bayesian classifiers (Xu et al., 2017), and support vector machines (SVM) (Schuller et al., 2010), are the main focus of research used to classify speech emotions. Most speech emotion recognition techniques use low-level descriptors (LLDs) generated from speech signals as input features, such as pitch, voice quality features, and MFCCs (Anagnostopoulos et al., 2015).

Lately, a growing number of academics have begun processing raw audio data to create Spectrograms or Mel-spectrograms, which represent low-level features (Zhao et al., 2019; Satt et al., 2017). Then, using several convolutional layers, the spectrograms are transformed into a potential time series before being passed through a recurrent layer.

In this work, we aim to build a robust speech emotion recognition system based on a dual-stream CNN-Transformer fusion network (DS-CTFN) that effectively extracts maximum feature information in terms of temporal and spatial views from raw audio data by exploring both MFCCs and Mel-spectrograms in parallel.

An overview of the main contributions of the proposed SER system is listed below:

1. We properly examined and researched the literature on speech signal processing, applied our own experiments, and established that there is complementary information between the MFCCs and Mel-spectrograms, included in the temporal and spatial feature representation. These two features both provide information that is necessary for emotion classification. The dual-stream parallel architecture required to process audio signals and learn high-level feature representations is not present in the current SER systems. Therefore we propose a novel dual-level architecture that consists of two distinct parallel CNN-Transformer networks that process the MFCCs/Mel-spectrograms and extracts the Spectral and Temporal Features separately yet concurrently trained. We found that our new dual-stream model offers a significant increase in accuracy.
2. We introduce two unique light-weight CNN structures that process MFCCs and Mel-spectrograms inputs. These CNN structures capture different feature maps through the use of three parallel paths with varying filter sizes. This enables deep convolution blocks to extract high-level features while maintaining adequate sparsity.
3. We tested our framework and it has shown outstanding results. It outperforms the best-known state-of-the-art models with an accuracy of 97.64%, 99.42%, and 97.53% on RAVDESS, TESS, and EMO-DB datasets respectively.

The rest of this article is organized into sections as described next.

Sect. 2 discusses the SER-related work. Section 3 describes in depth the proposed framework and its primary components. Section 4 demonstrates the suggested SER system's experimental setup and practical findings. Section 5 illustrates the discussion and comparative analysis. The last section is the conclusion and future research direction.

2 Related work

In literature, digital signal processing has been an active research field over the last decade (Huang et al., 2015; Mao et al., 2017; Kumaran et al., 2021). Many researchers have explored diverse methods in this field for building efficient speech emotion recognition (SER) systems. SER has the potential to be exploited in a broad variety of smart products, particularly intelligent conversation technologies and voice assistant systems.

The SER systems can be defined as either unimodal or multimodal. Unimodal data approaches focus entirely on speech signals, whereas multimodal data approaches include both audio signals and lexical information, in addition to visual data in some settings. Multimodal methods outperform unimodal methods by 6–7%, which is not strange considering that they incorporate more data inputs.

Unimodal techniques commonly extract high-level features from the raw audio signals, such as MFCC features, and then input the outputs into a recurrent neural network (RNN) (Mirsamadi et al., 2017). Recently, scientists have begun transforming audio input into spectrograms, which constitutes low-level acoustic data (Satt et al., 2017; Zhao et al., 2019). Then, using convolutional layers and a recurrent layer, the spectrograms are encoded into a latent time series representation. Some experts believe that textual data should also be taken into account because audio signal input alone is insufficient for obtaining reliable predictions. However, when fueled by different emotions, two utterances with similar lexical content may have totally different meanings. Therefore, focusing too much on textual information may be misleading and results in incorrect predictions. And the acoustic information included in the audio signal has a better potential for application in this task, for that reason, we focus solely on unimodal SER.

Deep learning techniques are gradually becoming the dominant option in this field and outperforming classical methods (Han et al., 2014; Stuhlsatz et al., 2011) Some of the main factors contributing to the success of deep learning techniques are the ability to extract complex features from input through a learning process. as the quantity of accessible data keeps expanding, as well as the remarkable improvement in computing capabilities (LeCun et al., 2015).

Since most deep-learning-based convolutional neural network (CNN) approaches are applied to 2D data, studies use deep spectrum methods for SER to extract high-level discriminative feature representations from speech spectrograms by using the CNNs. CNN models excel at extracting salient spatial features (Mao et al., 2014; Wang et al., 2019). Huang et al. (2014) was the first to use a CNN to extract the optimal emotion features from the speech signals, and they obtained remarkable results for SER on different datasets. Lately, fully convolutional networks (FCNs) have been introduced as a new version of CNN. FCNs handle inputs of varying sizes and have obtained good accuracy in time-series tasks (Naqvi et al., 2020). In contrast, FCN models are unable to extract salient temporal features. Lee and Tashev (2015) used a recurrent neural network (RNN) to model long-range temporal dependencies across speech frames. The recurrent neural network (RNN) and the long short-term memory(LSTM) provide state-of-art results in learning temporal dependency in time series (Karim et al., 2019; Zeng & Xiao, 2019). RNN-LSTM systems are frequently employed in SER tasks (Tao & Liu, 2018). Zhao et al. (2019) introduces a pairing of CNNs and RNNs-LSTMs that learns Spatio-temporal features and increased the accuracy of the previous SER systems.

SER networks have acquired potential performance increase with the introduction of Transformers (Vaswani et al., 2017), which have been regarded as a replacement for RNN-LSTM. Because traditional Recurrent Neural Networks are incapable of capturing temporal dependency across long distances. Instead, the Transformer employs the self-attention approach to attain rapid parallelism, which directly accesses global information to overcome the sequence distance constraint and extracts the most emotional semantic features from the input sequence. Additionally, Transformers are also non-recurrent, which reduces the training time significantly.

In this study, the goal is to improve the state-of-the-art performance of speech emotion recognition by exploring maximum feature information from raw audio data by encoding both MFCCs and Mel-spectrograms and effectively extracting both spatial and temporal data. Also, reduce the cost complexity by designing a light-weight structure and employing an effective parallel execution.

3 Methodology

The overall architecture of the proposed dual-level model is illustrated in Fig. 2.

This structure necessitates as inputs two distinct features: MFCCs, and Mel-spectrograms, which are fed into two parallel networks. Each network combines a CNN and a Transformer that extracts spatial features and temporal features information respectively from MFCCs and Mel-spectrogram inputs. Finally, the completed feature embeddings are then concatenated, used as a fully-connected layer input, and passed through the emotion classifier to obtain the SER class probabilities.

This section is subdivided into three segments, we will first introduce the data pre-processing phase, Next, we will explain the proposed two-stream architecture representation, and lastly define our emotion classifier.

3.1 Pre-processing

Mel-Frequency Cepstral Coefficients (MFCC) and Mel-spectrogram features of speech are time-frequency features. The computational process for Mel-spectrograms and MFCCs is similar, both are calculated through the same procedure, where MFCCs can be obtained by performing further steps. Figure 1 outlines the different steps involved in generating MFCC and log Mel-spectrogram features from an audio signal.

We generated the MFCC and Mel-spectrogram features using the “librosa” python package (McFee et al., 2015) and by applying the following parameters:

- *Framing* Human speech can fluctuate in length, Therefore framing is required to set the same input size. Since sound signals are non-stationary, it is reasonable to extract all of their spectral characteristics as a whole. Instead, we assume that the signal frequencies are stationary for a relatively little period of time.
- *Windowing* The signal is divided into frames, and then a window function is applied (Bingol & Aydogmus, 2020). windowing helps minimize the discontinuities of the audio signal by tapering each frame at the beginning and end edges. Frames are filtered by multiplying each frame by a specific window function that has the same size as the frame.
- *FFT* The Fast Fourier Transform (FFT) is a fast commonly used algorithm for the discrete Fourier transform (DFT) (Afrillia et al., 2017). It is employed to generate the frequency spectrum of each frame. to generate the frequency spectrum for each frame, we utilize the formula (1).

$$X_k = \sum_{k=0}^{N-1} X_k e^{\frac{-j2\pi kn}{N}} \tag{1}$$

- *Mel Filter Bank* Mel-scale is designed to be more discriminative at lower frequencies and less discriminative

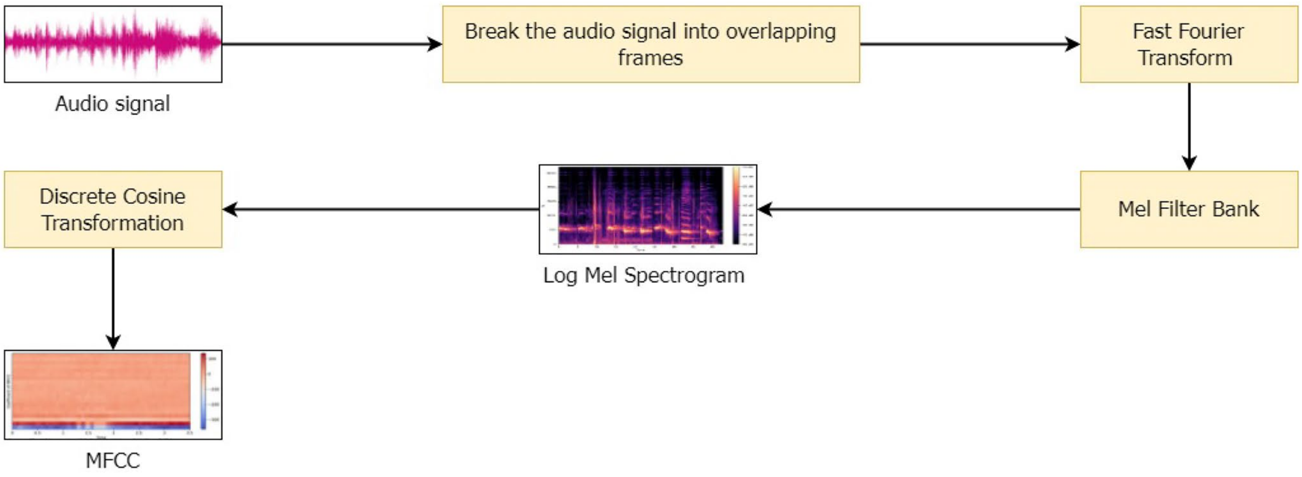

Audio signal
Break the audio signal into overlapping frames
Fast Fourier Transform
Mel Filter Bank
Log Mel Spectrogram
Discrete Cosine Transformation
MFCC

Fig. 1 MFFCs/Mel-spectrogram generation process

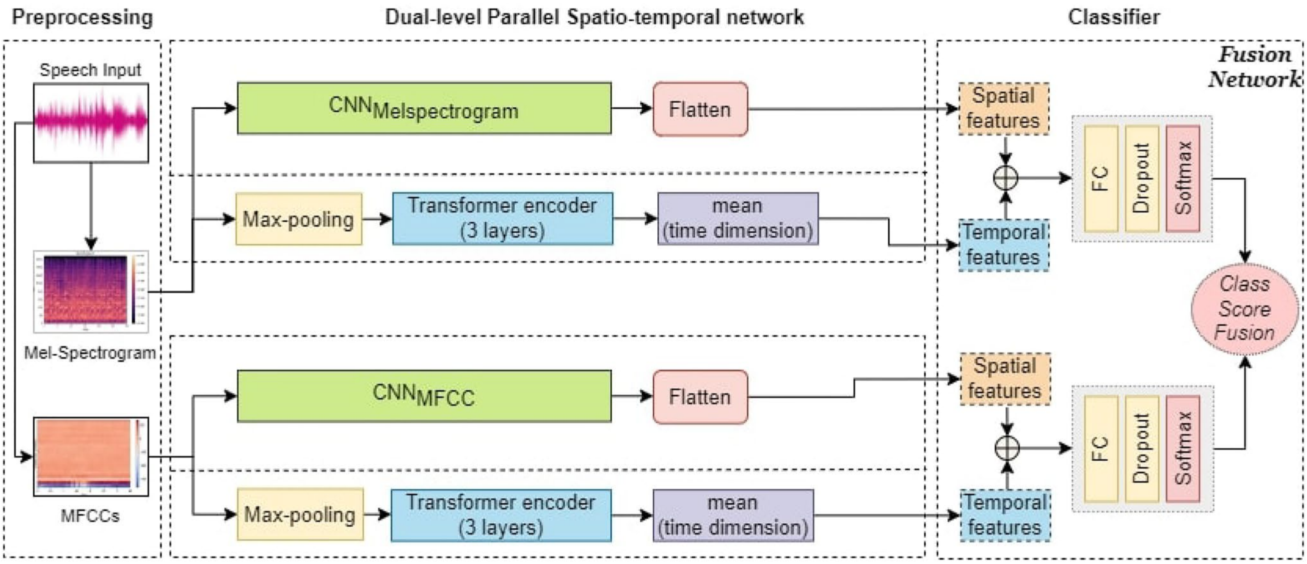


Fig. 2 Overall representation of the novel Dual-Stream CNN-Transformer Fusion Network for SER

at higher frequencies, imitating the non-linear perception of sound by the human ear. We produce Mel-spectrograms by transforming spectrograms from frequencies to mel scale. We use the formula(2) to calculate the mel frequency from the hertz frequency.

$$Mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \tag{2}$$

- *Logarithm* The human ear doesn't recognize sounds on a linear scale (Ismail et al., 2014), therefore the following phase is to calculate the log Mel-spectrogram by computing the logarithm of the Mel-spectrogram.
- *DCT (Discrete Cosine Transform)* We produce MFCC(Mel Frequency Cepstral Coefficients) features (Rahaman et al., 2019) by applying (DCT) to log Mel-spectrogram, and the number of MFCCs is 40.

For feature extraction, we solely use Mel-spectrograms and MFCCs. As we found through the experiments that taking Mel-spectrograms + MFCCs together provides the highest accuracy in this network, therefore we avoid increasing the complexity.

3.2 Proposed two-stream architecture

The following is a detailed description of our dual-stream CNN-Transformer fusion network (DS-CTFN). The main objective of this model is to obtain the optimal feature embedding representation in order to improve the performance of speech emotion recognition systems. An MFCC stream and a Mel-spectrogram stream comprise the dual-stream architecture. The model extracts maximum feature information in terms of temporal and spatial views from raw audio data by exploring both streams in parallel. This method has the potential to increase the performance of a wide range of applications, including speech recognition, speaker identification, and audio classification.

3.2.1 Spatial features extracted by CNN

Convolutional neural network models (CNNs or ConvNets) have become the state-of-the-art for solving visual imagery problems such as image recognition and have found success in various tasks involving speech/emotion, including automatic speech recognition (ASR), speech generation, and emotion recognition (Zhang et al., 2017; Li et al., 2019; Zayene et al., 2020). However, the challenge of variance in audio signals, consisting of a collection of speakers, diverse environmental settings, among others, limits the accuracy of emotion detection.

CNNs enable translation that is temporal and spatial independent, meaning that the network can still identify the class of an input even if it undergoes transformations. The pooling operation produces translational invariance, minimizing the effect of speech signal variability. Thus, applying the concept of a convolutional neural network in the acoustic modeling of speech emotion recognition allows us to overcome the challenge of audio signal variance.

In this work, we regard the Mel Frequency Cepstral Coefficients (MFCCs) and Mel-spectrograms generated from the audio data as images. Therefore, we utilize a deep convolution network, commonly employed in image analysis, for feature extraction. A convolutional neural network is a mathematical structure (Ciresan et al., 2011). The architecture is a multi-layered feed-forward neural network built by piling multiple hidden layers on top of each other in series, which enables it to learn hierarchical features.

Features are extracted through the convolution layers and pooling layers, while fully connected layers convey the extracted features to the final output for classification. The filters in the convolution layer with a specific height and width are shifted from left to right across the input picture. Feature map values are computed using the equation:

$$M[i,j] = (R * w)[i,j] = \sum_{x}\sum_{y} w[x,y]R[i-x,j-y] \tag{3}$$

where M represents the output feature map, R represents the input image, and w represents a kernel of size (x, y). The indexes of the rows and columns of the result matrix are denoted by i and j, respectively. This equation calculates the convolution between the input image and the filter. The filter is shifted across the input image, and at each position, the element-wise product between the filter and the corresponding part of the input image is computed, and then all the products are summed to get the value at the output feature map.

Overall, the combination of convolutional and pooling layers in CNNs enables effective feature extraction from images, which can also be applied to audio data by treating the MFCCs/Mel-spectrograms as images.

The proposed structure of our CNN has two distinctive parts:

- *Stage I* In this stage, the input MFCC/Mel-spectrogram signals undergo convolution operations with three parallel sets of varying kernel sizes, as depicted in Fig. 3. The primary objective of applying this parallel set of convolution layers is to extract as many distinct feature embedding combinations as possible. Previous research works (Araujo et al., 2019; Aftab et al., 2022) have demonstrated a strong relationship between kernel size and prediction accuracy. It has been shown that expanding the receptive field can enhance the classification accuracy of the model. As such, to increase the receptive field of the model without significantly increasing the number of trainable parameters, we employed convolution kernels of sizes (11×1), (1×11), and (3×3) as shown in Fig. 4.

  Moreover, deeper networks with more layers lead to larger receptive fields since each added layer expands the receptive field by the kernel size (Luo et al., 2016). However, deep networks tend to overfit the training data due to the increased number of trainable parameters. By using parallel convolution layers with different kernel sizes, we can extract a more diverse set of features while reducing the number of parameters and the computational cost of the model.

  Another benefit of using the two 1-D kernels (11×1) and (1×11) is that each focuses more on the shifted dimension. Finally, the extracted features from the three parallel paths are concatenated and fed into the Stage II of the model.

  In summary, the use of three parallel sets of convolution layers with different kernel sizes in Stage I allows us to extract a diverse set of features from the input signals while reducing the number of parameters and computational cost of the model.

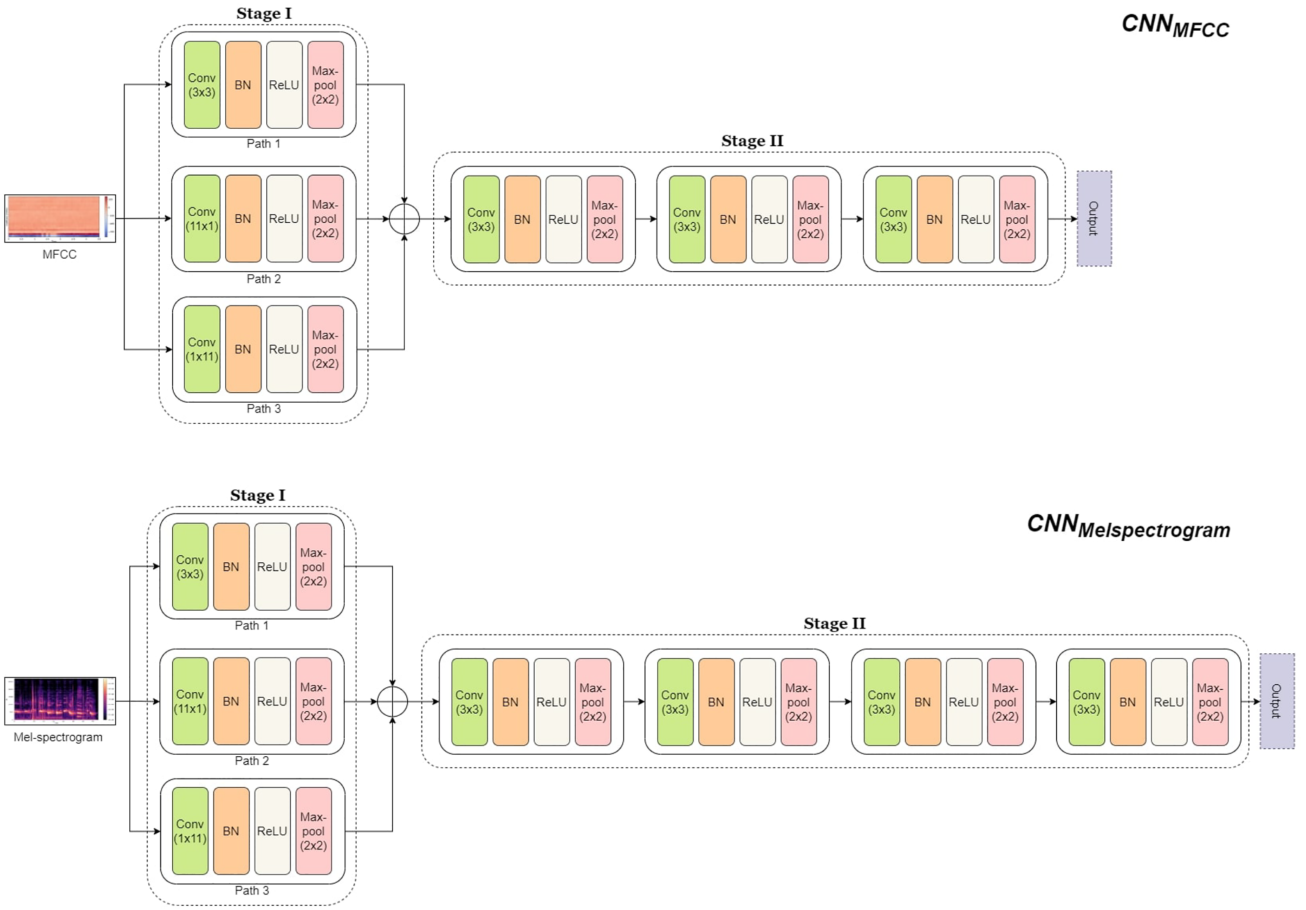


Fig. 3 The detailed structure of CNN$_{\text{MFCC}}$ and CNN$_{\text{Mel-spectrogram}}$

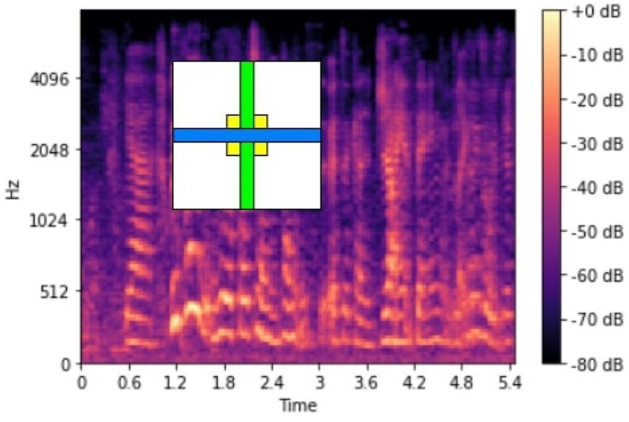


Fig. 4 The resultant receptive fields (white rectangle) of three parallel paths (yellow, green, and blue rectangles). (Color figure online)

- *Stage II* In this stage, a Convolutional Neural Network (CNN) is employed to further process the extracted features from Stage I. The CNN consists of four sequences of layers, as illustrated in Fig. 3. However, since the dimension of MFCCs is much smaller than Mel-spectrograms due to the difference in the number of Mel-frequency bands/coefficients, only three sequences are used for CNN$_{\text{MFCC}}$.

Each sequence in the CNN comprises a convolutional layer, a batch normalization layer (BN), a rectified linear unit (ReLU), and a max-pooling layer. The convolutional layer extracts feature maps by applying a set of filters to the input feature maps. The batch normalization layer normalizes the activations of the previous layer to improve the training speed and performance of the network. The rectified linear unit (ReLU) is an activation function that introduces non-linearity into the network, allowing it to learn more complex features.

The max-pooling layer downsizes the feature maps by selecting the maximum value within a given region. This operation reduces the dimensionality of the feature maps, thereby reducing the number of parameters needed for computation, and preventing overfitting. The resulting output of the max-pooling layer is a feature map that includes the most relevant features from the prior feature map.

In summary, Stage II of the model involves the use of a CNN with multiple sequences of layers to process the extracted features from Stage I. The use of max-pooling allows for the downsampling of the feature maps, reducing the computational cost of the model and preventing overfitting. This stage is crucial for extracting high-level features that can be used for classification in the subsequent stage of the model.

3.2.2 Temporal features extracted by transformers

For the second parallel stream, we adopt the Transformer-Encoder that was first presented in "Attention is All You Need" (Vaswani et al., 2017). The intended purpose is to recognize the frequency distributions of various emotions based on the general structure of Mel-spectrogram/MFCCs for each emotion. Let $X \in \mathbb{R}^{T \times F}$ denote the input Mel-spectrogram/MFCC feature map, where T is the number of time frames and F is the number of Mel-frequency bands/coefficients.

We start by max-pooling the initial input to obtain a feature map $\hat{X} \in \mathbb{R}^{1 \times F}$, which reduces the computation cost.

Next, we apply 3 multi-head self-attention layers to the pooled feature map, which can be expressed as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \dots, head_h)W^O \tag{4}$$

where

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

and

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Here, Q, K, and V are the query, key, and value matrices respectively, and W_i^Q, W_i^K, and W_i^V are the learnable weight matrices for the i^{th} attention head. h is the number of attention heads, We use $h = 8$ attention heads. Concat is the concatenation operation. The output of each attention head is concatenated and then projected back to the original feature dimensionality using the output weight matrix W^O.

We use the ReLU activation function $\text{ReLU}(x) = \max(0, x)$ to reduce the computation time and avoid saturation. We also apply a 40% dropout regularization to avoid overfitting.

Lastly, to get a one-dimensional feature vector, we compute the mean value on the second dimension, which can be expressed as follows:

$$\text{MeanPool}(X) = \frac{1}{F} \sum_{i=1}^{F} X[:, i] \tag{7}$$

The motivation for choosing Transformers over LSTM/RNNs is the fact that multi-head self-attention layers in the Transformer allow the network to consider several prior time steps as the emotions can be present all across the sequence, while LSTM/RNNs only recognize frequency variations according to the adjacent time steps. This makes the Transformer more effective in capturing long-term dependencies in the temporal domain.

3.3 Classifier

The classifier in our model is a fusion network. This module is responsible for combining the features extracted from the temporal and spatial streams, and producing the final probability distribution across the emotion classes. The classifier structure is illustrated in Fig. 2.

For each stream, we first flatten the output layer of the CNN into a one-dimensional vector and concatenate it with the output layer of the Transformer Encoder to integrate the two parallel parts (Temporal+Spatial). The concatenated feature vector is denoted by h_{concat} and can be expressed as:

$$h_{concat} = [h_{CNN}; h_{Transformer}] \tag{8}$$

where h_{CNN} and $h_{Transformer}$ are the output feature maps of the CNN and Transformer-Encoder, respectively.

The concatenated feature vector h_{concat} is then passed through a dense layer followed by a softmax activation function to obtain the probability distribution over the emotion classes. Let W and b denote the weight matrix and bias vector of the dense layer, respectively. Then, the class probabilities p_i for each emotion class i are given by:

$$z = W \cdot h_{concat} + b \tag{9}$$

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{10}$$

where C is the total number of emotion classes.

Next, we employ a late fusion strategy, where the outputs of both the MFCC stream and Mel-spectrogram stream are fused using the class probabilities. Both streams contribute equally to the resulting output producing the final classification probability. The final probability distribution over the emotion classes p_{final} is obtained by averaging the class probabilities obtained from the two streams:

$$p_{final} = \frac{p_{MFCC} + p_{Mel}}{2} \tag{11}$$

where p_{MFCC} and p_{Mel} are the class probabilities obtained from the MFCC stream and Mel-spectrogram stream, respectively.

4 Experimental settings and results

In this section, to demonstrate the significance and effectiveness of our system, we evaluated our SER model on the RAVDESS, TESS, and EMO-DB speech datasets using MFCCs and Mel-spectrograms. We compared the performance and results with other recent SER baseline models. Several experiments were conducted. the datasets used, the accuracy metrics and the detailed results are discussed in the upcoming subsections.

4.1 Dataset description

We use three datasets to assess the effectiveness of our model: RAVDESS dataset (Livingstone & Russo, 2018), TESS dataset (Dupuis & Pichora-Fuller, 2010), and EMO-DB dataset (Burkhardt et al., 2005).

The following is the description of each dataset.

RAVDESS The RAVDESS database(Ryerson Audiovisual Database of Emotional Speech and Songs) (Livingstone & Russo, 2018) is a new emotion corpus frequently utilized to train models for emotion recognition of speakers in emotional songs and emotional speech. It is an acted dataset recorded in the British English language. The RAVDESS corpus contains (8) emotions recorded by 24 professional actors (12 male, and 12 female). The RAVDESS dataset has a total of 1440 utterances recorded with a sample rate of 48 kHz and the average time of each utterance is 3.5 s.

EMO-DB The Berlin emotion database Emo-DB (Burkhardt et al., 2005) is an emotion corpus that contains a total of 535 utterances recorded in the German language by 10 professional actors (5 male and 5 female). The average length of utterances is between 3 to 5 s recorded with a sample rate of 16 kHz. The EMO-DB dataset is widely used for speech emotion recognition research due to the clarity of the recordings which improves the emotion recognition rate.

TESS The Toronto Emotional Speech Set(TESS) is an English-language Speech Emotion dataset produced at the University of Toronto in 2010 Dupuis and Pichora-Fuller (2010). The corpus contains 2800 audio recordings portraying speech in seven various emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. Two native English-speaking actresses (aged 26 and 64 years) were conveying the 200 target words. the sample rate of the recordings is 22,050 Hz and the average time of each utterance is 2.06 s.

Table 1 provides a summary of the emotion classes, numbers of utterances, and participation rates for three datasets, namely RAVDESS, EMO-DB, and TESS. The numbers of utterances indicate the number of audio recordings associated with each emotion class. The contribution rates indicate the percentage of utterances contributed by each emotion class to the overall dataset.

Table 1 The detailed description of the emotion classes in the RAVDESS, EMO-DB, and TESS datasets, including the numbers of utterances per class and the contribution rates

Emotion	RAVDESS	EMO-DB	TESS
Anger	192 (13.33%)	127 (23.74%)	400 (14.29%)
Calm	192 (13.33%)	–	–
Disgust	192 (13.33%)	46 (8.60%)	400 (14.29%)
Fear	192 (13.33%)	69 (12.90%)	400 (14.29%)
Happiness	192 (13.33%)	71 (13.27%)	400 (14.29%)
Neutral	96 (6.67%)	79 (14.77%)	400 (14.29%)
Sadness	192 (13.33%)	62 (11.59%)	400 (14.29%)
Surprise	192 (13.33%)	–	–
Boredom	–	81 (15.14%)	–
Pleasant S	–	–	400 (14.29%)

The datasets RAVDESS, EMO-DB, and TESS are often utilized in the development of speech emotion recognition (SER) models. However, as can be noticed in the above descriptions, there are certain disparities between these datasets that might affect the performance of the SER model. The RAVDESS dataset contains 7356 audio samples, comprising speech and song, performed by 24 professional actors (12 male and 12 female) in 8 different emotions. RAVDESS offers a balanced gender and emotion distribution, making it excellent for training and evaluating SER models. EMO-DB is a German database of emotional speech utterances delivered by 10 actors (5 male and 5 female) in 7 distinct emotions. This dataset only contains German speech, which limits its utility in constructing SER models for other languages. Also, EMO-DB has a smaller sample size than RAVDESS and TESS datasets. TESS is a database of 2800 audio recordings of emotional speech delivered by two female actresses. TESS offers 7 distinct emotions. However, due to the small number of speakers and the lack of male speakers, it may not be indicative of the wider population as a whole. As a result, the dataset used is determined by the specific objectives of the SER model development. Given its large size, balanced distribution, and availability of speech and song samples, RAVDESS is a favorable choice for creating SER models. Nevertheless, if the goal is to build a SER model for German speech, EMO-DB may be better. TESS can be employed as well, however, the small number of speakers might not represent the general population.

4.2 Evaluation metrics

We use the following three assessment measures to compare and evaluate performance:

1. *Unweighted average recall (UAR)* This is the unweighted average of the class-specific recalls and is defined as:

$$UAR = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i}$$

where n is the total number of emotion classes, TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i.

2. *Weighted accuracy (WA)* This is the overall classification accuracy of all test data and is defined as:

$$WA = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)}$$

3. *F1-score or F1-measure* This is the harmonic mean between precision and recall and is defined as:

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

In our case, UAR is the most important metric as the accuracy of each emotion class cannot be accurately represented by the overall accuracy (WA). The effectiveness of the model in dealing with unbalanced data, such as in the EMO-DB dataset, may be reflected more clearly with the Unweighted average recall (UAR) metric.

4.3 Experimental setup and configuration

The proposed SER network comprises a new architecture and achieves high results in emotion recognition using MFCCs and Melspectrogram features. We use three benchmark datasets to test the performance of our model: the RAVDESS dataset, the TESS dataset, and the EMO-DB dataset. The recommended datasets were randomly divided into an 80:20 ratio. 20% of the data was utilized for the testing phase, while 80% was used for the training phase.

In the training phase, we adjusted the model with different parameters to select the optimal set of parameters for SER. We ran many experiments with different batch sizes, learning rates, and the number of Transformer layers to select the optimal configuration. We chose a learning rate of 1e−4, a batch size of 32, and set the number of epochs to 250. The model training and the evaluation presented in the article were performed in a uniform environment, which consisted of an NVIDIA K80 GPU with 12GB of RAM, a 2-core Intel Xeon CPU running at 2.30GHz, and 12GB of memory. We implemented the proposed SER system using the python deep learning library PyTorch (Paszke et al., 2019). We used the Stochastic Gradient Descent(SGD) optimizer and cross-entropy loss as our objective function. Every 100 training steps, we apply an early stop by checking the performance accuracy on the validation set.

For the CNN modules used to process the mel-spectrograms and MFCCs, the detailed structure is presented in Fig. 3.

The training results demonstrate the significant impact of the fusion network on creating an effective SER model. The graphical charts display the training and validation accuracies and losses for each of the recommended datasets (see Figs. 5, 6).

4.4 Model recognition analysis

In order to evaluate the adaptability of our model for the real world, we assessed the prediction accuracy of the new SER system using unseen data that were partitioned before the model was trained. For model testing, we used a variety of classification reports and confusion matrices.

The precision, recall, and F1-measure per class, as well as the model's overall prediction accuracy, are included in the classification report. The classification performance is reflected by the confusion matrix, where the diagonal and

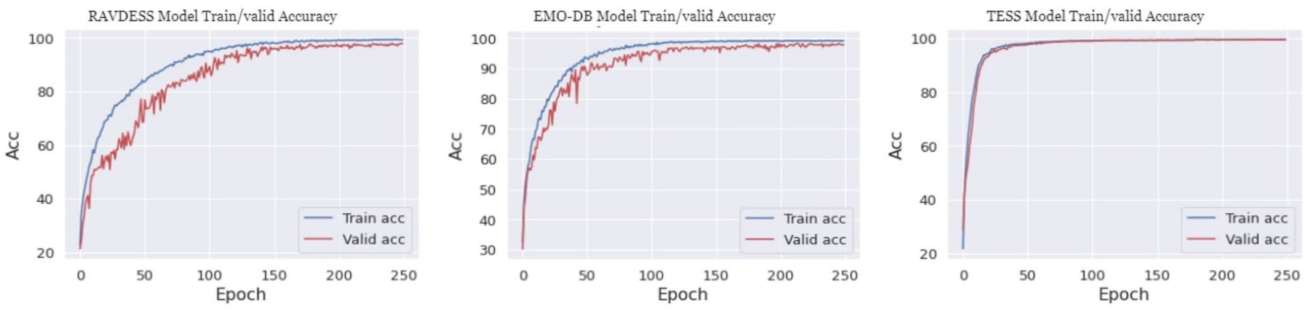


Fig. 5 The training performance of our DS-CTFN model. The three charts represent the model training/validation accuracies for the three datasets

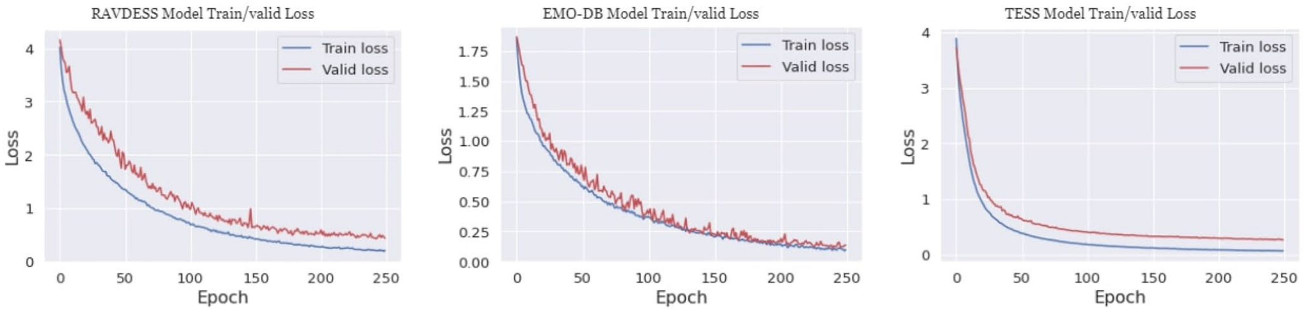


Fig. 6 The training performance of our DS-CTFN model. The three charts represent the model training/validation losses for the three datasets

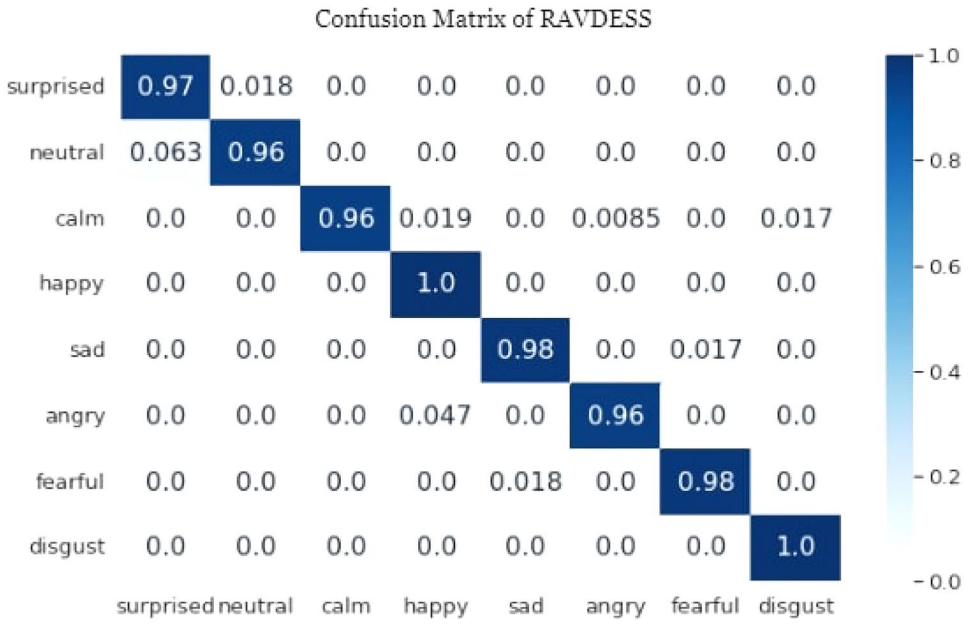
off-diagonal values correspond to the correct and incorrect emotional predictions, respectively. We examined the model on various emotions and got the confusion matrices and classification reports for the RAVDESS, TESS, and EMO datasets.

The classification report of the model on the three benchmarks is presented in Table 2. We evaluate and examine the performance of the model for each emotion and measure its accuracy. We also calculate the weighted and unweighted accuracy for each dataset in order to evaluate the model and compare it to other state-of-the-art (SOTA) methods. For each emotion class, the report displays the precision, recall, and F1-measure using weighted and unweighted averages, however, it is missing information on confusion between the classes. Therefore, for in-depth analysis, we employ the confusion matrix, which highlights the discrepancy between true classes and predicted classes and also displays their level of confusion with other emotions at related rows.

Figures 7, 8, and 9. Are the illustration of the computed confusion matrices with class-wise accuracies for RAVDESS, EMO-DB, and TESS datasets respectively. For the RAVDESS dataset, the happy and disgust classes had the perfect score of (1.0). For the EMO-DB dataset, the class neutral obtained the best result scoring (1.0). The least performing class was happy with a score of (0.95). For the TESS dataset, which is a less challenging dataset, most of the classes obtained the perfect score.

Overall, the classification report and confusion matrices provide a comprehensive evaluation of the performance of our proposed SER system, highlighting its strengths and areas for improvement.

Fig. 7 Confusion matrix of the proposed SER model using RAVDESS dataset with 97.64% average recognition accuracy


Confusion Matrix of RAVDESS

	surprised	neutral	calm	happy	sad	angry	fearful	disgust
surprised	0.97	0.018	0.0	0.0	0.0	0.0	0.0	0.0
neutral	0.063	0.96	0.0	0.0	0.0	0.0	0.0	0.0
calm	0.0	0.0	0.96	0.019	0.0	0.0085	0.0	0.017
happy	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
sad	0.0	0.0	0.0	0.0	0.98	0.0	0.017	0.0
angry	0.0	0.0	0.0	0.047	0.0	0.96	0.0	0.0
fearful	0.0	0.0	0.0	0.0	0.018	0.0	0.98	0.0
disgust	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

1.0
0.8
0.6
0.4
0.2
0.0

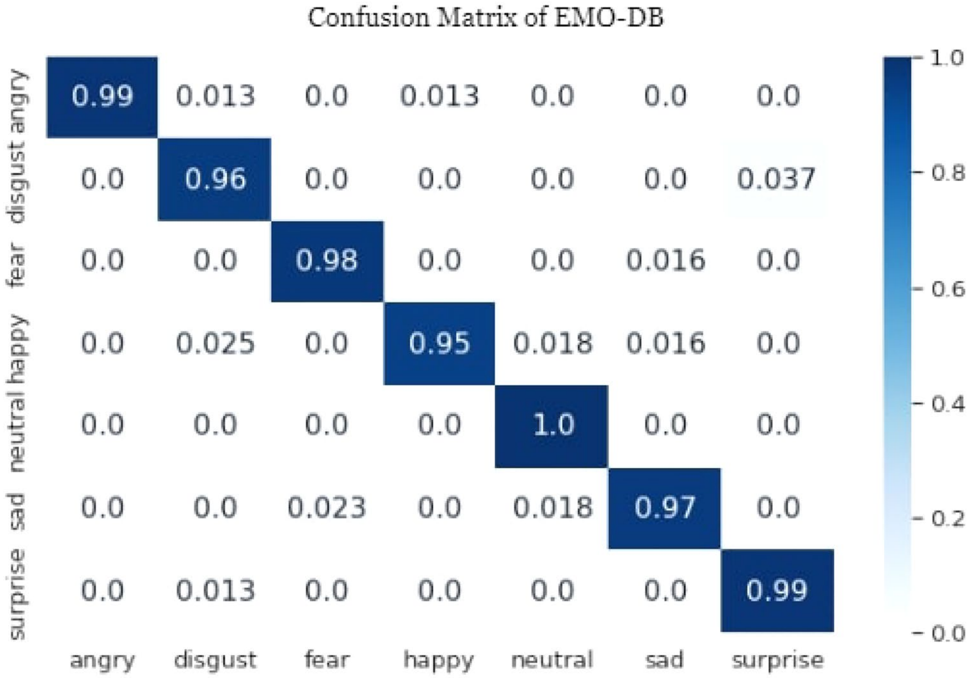


Fig. 8 Confusion matrix of the proposed SER model using EMO-DB dataset with 97.53% average recognition accuracy

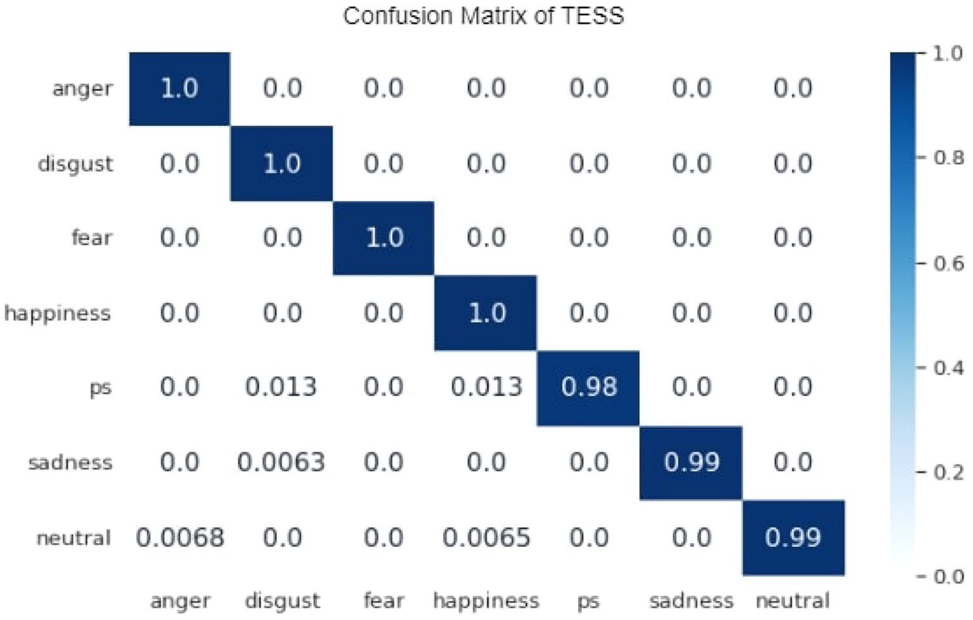


Fig. 9 Confusion matrix of the proposed SER model using TESS dataset with 99.42% average recognition accuracy

4.5 Model ablation

4.5.1 Baseline models

We selected six baseline models composed of different parts of the entire model in order to properly evaluate the impact of each component.

- *Base 1:* in experiment 1, we consider the sub-network part (CNN$_{\text{MFCC}}$+Transformer) that takes as input only MFCCs. The stageI of CNN$_{\text{MFCC}}$ here is constituted by only one path of (3×3) kernel.
- *Base 2:* experiment 2 is similar to experiment 1, where we consider the other sub-network part (CNN$_{\text{Mel-spectrogram}}$+Transformer) that takes as input only Mel-spectrograms. The stage I of the CNN$_{\text{Mel-spectrogram}}$ here is constituted by only one path of (3×3) kernel.
- *Base 3:* experiment 3 is the combination of Base1+Base2, we consider the dual stream full network that takes as input both MFCCs + Mel-spectrograms, However, the

Table 2 Classification report of the DS-CTFN model on RAVDESS, TESS, and EMO-DB datasets

RAVDESS				TESS				EMO-DB			
Emotion	Precision	Recall	F1	Emotion	Precision	Recall	F1	Emotion	Precision	Recall	F1
Surprised	0.938	0.968	0.953	Anger	0.993	1.000	0.997	Anger	1.000	0.985	0.992
Neutral	0.981	0.964	0.972	Disgust	0.981	1.000	0.991	Boredom	0.951	0.963	0.957
Calm	1.000	0.957	0.978	Fear	1.000	1.000	1.000	Disgust	0.977	0.977	0.977
Happy	0.938	1.000	0.968	Happy	0.981	1.000	0.990	Fear	0.986	0.947	0.966
Sad	0.982	0.982	0.982	PS	1.000	0.977	0.988	Happy	0.966	1.000	0.966
Anger	0.991	0.958	0.974	Sad	1.000	0.994	0.997	Sad	0.968	0.968	0.968
Fear	0.983	0.983	0.983	Neutral	1.000	0.989	0.995	Neutral	0.964	0.988	0.976
Disgust	0.983	1.000	0.992								
Average	0.975	0.976	0.975		0.994	0.994	0.994		0.973	0.975	0.974
Weighted avg.	0.977	0.977	0.977		0.994	0.994	0.994		0.976	0.976	0.976
Accuracy	97.64%			99.42%				97.53%			

stage I of both CNN_{MFCC} and $\text{CNN}_{\text{Mel-spectrogram}}$ are constituted by only one path of (3×3) kernel.

- *Base 4:* experiment 4, is similar to experiment 1, we consider the sub-network part (CNN_{MFCC}+Transformer) that takes as input only MFCC features. However, stage I of CNN_{MFCC} here is constituted by 3 parallel paths with different kernel sizes: (3×3), (11×1), and (1×11).
- *Base 5:* experiment 5, is similar to experiment 4, where we consider the other sub-network part ($\text{CNN}_{\text{Mel-spectrogram}}$+Transformer) that takes as input only Mel-spectrogram.
- *Base 6:* experiment 6 is our proposed model, it is the combination of Base 4 and Base 5. The dual stream full network takes as input both MFCCs + Mel-spectrograms and processes them all in parallel.

4.5.2 Results analysis

We present a detailed analysis of the results obtained from various network configurations of the DS-CTFN model, as shown in Table 3. The models were evaluated based on three

Table 3 The DS-CTFN model performance compared to different network variants on RAVDESS, TESS, and EMO-DB datasets

Database	Baseline	# Param	Input	WA	UAR	F1
RAVDESS	Base 1	232,594	MFCCs	0.777	0.762	0.766
	Base 2	717,640	Mel-spectrograms	0.944	0.944	0.943
	Base 3	953,338	MFCCs + Mel-spectrograms	0.970	0.970	0.969
	Base 4	251,474	MFCCs	0.806	0.799	0.803
	Base 5	736,520	Mel-spectrograms	0.953	0.952	0.951
	Base 6	961,178	MFCCs + Mel-spectrograms	**0.977**	**0.976**	**0.975**
TESS	Base 1	232,337	MFCCs	0.934	0.934	0.934
	Base 2	716,935	Mel-spectrograms	0.975	0.975	0.975
	Base 3	952,376	MFCCs + Mel-spectrograms	0.982	0.982	0.982
	Base 4	251,217	MFCCs	0.976	0.976	0.976
	Base 5	735,815	Mel-spectrograms	0.978	0.978	0.978
	Base 6	953,672	MFCCs + Mel-spectrograms	**0.994**	**0.994**	**0.994**
EMO-DB	Base 1	232,757	MFCCs	0.834	0.812	0.819
	Base 2	717,831	Mel-spectrograms	0.938	0.936	0.934
	Base 3	953,692	MFCCs + Mel-spectrograms	0.966	0.964	0.963
	Base 4	251,637	MFCCs	0.850	0.839	0.844
	Base 5	736,711	Mel-spectrograms	0.940	0.937	0.938
	Base 6	954,400	MFCC + Mel-spectrograms	**0.976**	**0.975**	**0.974**

The values indicated in bold represent the top-performing results for weighted accuracy (WA), unweighted average recall (UAR), and F1 score across each respective database different metrics: Weighted Accuracy (WA), Unweighted average recall(UAR), and F1-score.

For the RAVDESS database, the results show that Base 6 achieved the highest WA (0.977), UAR (0.976), and F1 score (0.975). It is interesting to note that Base 3 achieved higher performance than Base 1 and Base 4, which also used MFCCs as input. Moreover, Base 2 and Base 5 achieved a high performance using Mel-spectrograms as input. However, their performance is not as high as Base 6, which used both Mel-spectrograms and MFCCs as input.

For the TESS database, the results show that Base 6 had the highest WA, UAR, and F1 score among all the models, achieving a score of (0.994) for all the metrics. Again, Base 3 achieved higher performance than Base 1 and Base 4, which used MFCCs as input. Similarly, Base 2 and Base 5 achieved a high performance using Mel-spectrograms as input. However, their performance is not as high as Base 6, which used both Mel-spectrograms and MFCCs as input.

The findings from analyzing the EMO-DB database reveal that Base 6 achieved the highest WA (0.934), UAR (0.933), and F1 score (0.934). Comparatively, Base 3 outperformed Base 1 and Base 4, which relied solely on MFCCs as input. Moreover, Base 2 and Base 5 also demonstrated good performance by using Mel-spectrograms as input. Nonetheless, their performance was lower than that of Base 6, which utilized both Mel-spectrograms and MFCCs as input.

First, we compare the performance of configurations with similar basic structures but different inputs, specifically Base 1 and Base 2, compared to Base 4 and Base 5. The former pair uses MFCCs as input, while the latter pair uses Mel-spectrograms. We observe that the configuration that uses Mel-spectrograms input has a larger number of trainable parameters, which results in a more complex structure and higher training cost. However, it offers between 0.2 and 18.2% relative improvement in UAR compared to the configuration that takes only MFCCs. We attribute this improvement to the fact that Mel-spectrograms are less processed and contain more details and information about the emotions in the utterance that the model can learn from.

Next, we evaluate the performance of configurations that use a dual-level structure, which takes both MFCCs and Mel-spectrograms as inputs, namely Base 3 and Base 6. We compare the results of these two baseline models with the rest of the configurations that take only single inputs. Interestingly, we observe a 01.6% to 21.4% relative improvement in UAR for the dual-level structures, despite being more complex. Moreover, the parallel structure significantly lowers the cost of training and makes it negligible compared to the important improvement in performance scores. The dual-level model DS-CTFN fully explores the complementary information between the MFCCs and Mel-spectrograms, resulting in outperforming Bases 1 to 5 and confirming the efficacy of the DS-CTFN dual-level architecture.

We also investigate the impact of the number of paths in the first stage of $\mathrm{CNN}_{\mathrm{MFCC}}$ and $\mathrm{CNN}_{\text{Mel-spectrogram}}$ in Base 1–3 and Base 4–6, respectively. Configurations in Base 1–3 have Stage I composed of only one path, while Base 4–6 have Stage I composed of 3 parallel paths. We observe that the triple path structures expand the receptive field and enhance classification accuracy results by 0.1% to 4.2%, yet there is no significant increase in the number of trainable parameters and the computation cost. This highlights the positive impact of the parallel strategy.

Finally, we evaluate the impact of feature normalization on the overall recognition accuracy of the DS-CTFN model, as shown in Table 4. We observe that feature normalization enhances the overall recognition accuracy for RAVDESS by 2.52%, for EMO-DB by 1.31%, and for TESS by 0.64% compared to un-normalized features. This suggests that normalization helps the model to learn more discriminative features by reducing the effect of varying scales and ranges of feature values.

Overall, our results demonstrate that the DS-CTFN model's efficacy lies in its ability to leverage the complementary information between MFCCs and Mel-spectrograms and to utilize a parallel strategy in the first stage of the CNN to expand the receptive field without adding significant computational cost. Additionally, the normalization of features can further enhance the recognition accuracy of the DS-CTFN model.

4.5.3 Comparison with the state-of-the-art methods

Table 5 displays the comparative performance results of the proposed SER system with the best existing state-of-the-art (SOTA) baseline SER models tested on the RAVDESS, TESS, and EMO-DB datasets. Our model surpasses state-of-the-art results. We present and compare the performance accuracy of our proposed model, for RAVDESS, the emotion recognition accuracy of our model is at least 0.34% higher than the accuracy of the Gumelar et al. (2022) model. For the TESS dataset, the emotion recognition accuracy score of our model is at least 2.32% higher than the accuracy of the Choudhary et al. (2022) model. For the EMO-DB

Table 4 Model recognition accuracy with normalized and un-normalized values for speech emotion recognition on RAVDESS, TESS, and EMO-DB datasets

	RAVDESS (%)	TESS (%)	EMO-DB (%)
Proposed model + un-normalized features	95.12	98.78	96.22
Proposed model + normalized features	97.64	99.42	97.53

Table 5 The performance of the DS-CTFN model compared to best existing state-of-the-art models on RAVDESS, TESS, and EMO-DB datasets

RAVDESS			TESS			EMO-DB		
References	Year	Accuracy (%)	References	Year	Accuracy (%)	References	Year	Accuracy (%)
Zeng et al. (2019)	2019	64.48	Dupuis and Pichora-Fuller (2011)	2011	82.00	Guo et al. (2019)	2019	84.49
Jalal et al. (2019)	2019	69.40	Praseetha and Vadivel (2018)	2018	95.82	Meng et al. (2019)	2019	88.99
Bhavan et al. (2019)	2019	75.79	Huang and Bao (2019)	2019	85.00	Chen et al. (2018)	2018	82.82
Zamil et al. (2019)	2019	67.14	Choudhary et al. (2022)	2022	97.10	Badshah et al. (2019)	2019	80.79
Issa et al. (2020)	2020	71.61				Jiang et al. (2019)	2019	84.53
Kwon (2019)	2020	79.01				Li et al. (2021)	2021	82.10
Sajjad et al. (2020)	2020	77.01				Sajjad et al. (2020)	2020	85.57
Assunção et al. (2020)	2020	71.60				Anvarjon and Kwon (2020)	2020	89.02
Jason and Kumar (2020)	2019	80.21				Kwon (2021)	2020	90.01
Han et al. (2021)	2021	80.89				Mustaqeem (2021)	2021	91.11
Atila and Şengür (2021)	2021	96.18				Kwon (2021)	2021	95.00
Gong et al. (2021)	2021	97.00						
Gumelar et al. (2022)	2022	97.30						
Proposed method	2022	97.64	Proposed method	2022	99.42	Proposed method	2022	97.53

dataset, the emotion recognition accuracy of our model is at least 2.53% higher than the accuracy of the Kwon (2021) model. These results indicate the effectiveness of our proposed model.

5 Conclusion

In this study, we have presented an innovative and robust speech emotion recognition system based on a dual-stream CNN-Transformer fusion network that effectively extracts maximum feature information from raw audio data by exploring both MFCCs and Mel-spectrograms in parallel. Our approach has shown to be quite efficient. It outperforms the best-known state-of-the-art (SOTA) models with an accuracy of 97.64%, 99.42%, and 97.53% on RAVDESS, TESS, and EMO-DB datasets. Additionally, we examined how to explore and process raw audio data to properly extract the optimal set of features without increasing the complexity of the model, as well as how to effectively classify emotions using the extracted features.

We conducted several experiments on the speech emotion datasets to demonstrate the significant advantages of our proposed model. The findings show that our innovative architecture can learn to recognize emotional data features accurately.

In addition to our current results, there are other pathways for even further research in speech emotion recognition (SER). First of all, even though we have obtained encouraging results with the existing audio inputs, we intend to carry out a more thorough examination into the possibility of additional types of audio inputs, such as pitch, energy, and formants, to further enhance the accuracy and reliability of our model. Second, we would like to investigate more challenging datasets that reflect actual conversational contexts. The aim is to assess our model's adaptability to a variety of spoken interactions, these datasets will include complicated interaction scenarios between speakers, such as overlapped speech and background noise. Furthermore, we want to explore the usage of transfer learning strategies to train our model on a large dataset of non-emotional audio data before fine-tuning it and optimizing it on an emotional speech dataset. Using data collected from a bigger and more diversified collection of speech has the potential of improving the model's capacity to detect emotions. Lastly, another area of focus is the development of multimodal SER models capable of incorporating both auditory and visual inputs to increase emotion detection performance. Visual signals including facial expressions and body language convey vital information about a person's emotional state. By adding this input into our models, we may be able to attain even better performance in emotion recognition tasks.

Author contributions MT:Conceptualization, Data curation, Investigation, Methodology, Resources,Software, Validation, Visualization, Writing—original draft, Writing—reviewediting. LG: Formal analysis,

Validation, Writing—review editing.QM: Supervision, Formal analysis, Methodology, Validation,Writing—review editing.

Funding This work is supported in part by the Key Projects of the NationalNatural Science Foundation of China under Grant U1836220, the NationalNature Science Foundation of China of 62176106, and Jiangsu Province keyresearch and development plan (BE2020036).

Data and code availability The data and code that supports the results presented in this article are available upon request.

Declarations

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

References

Afrillia, Y., Mawengkang, H., Ramli, M., Fhonna, R. P., et al. (2017). Performance measurement of mel frequency ceptral coefficient (MFCC) method in learning system of al-qur'an based in nagham pattern recognition. *Journal of Physics, 930*, 012036.

Aftab, A., Morsali, A., Ghaemmaghami, S., & Champagne, B. (2022). Light-sernet: A lightweight fully convolutional neural network for speech emotion recognition. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6912–6916). IEEE.

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.

Anvarjon, T., & Kwon, S. (2020). Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors, 20*(18), 5212.

Araujo, A., Norris, W., & Sim, J. (2019). Computing receptive fields of convolutional neural networks. *Distill, 4*(11), 21.

Assunção, G., Menezes, P., & Perdigão, F. (2020). Speaker awareness for speech emotion recognition. *International Journal of Online and Biomedical Engineering, 16*(4), 15–22.

Atila, O., & Şengür, A. (2021). Attention guided 3d CNN-LSTM model for accurate speech based emotion recognition. *Applied Acoustics, 182*, 108260.

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., & Baik, S. W. (2019). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications, 78*(5), 5571–5589.

Bhavan, A., Chauhan, P., Shah, R. R., et al. (2019). Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems, 184*, 104886.

Bingol, M. C., & Aydogmus, O. (2020). Performing predefined tasks using the human-robot interaction on speech recognition for an industrial robot. *Engineering Applications of Artificial Intelligence, 95*, 103903.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of German emotional speech. *Interspeech, 5*, 1517–1520.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters, 25*(10), 1440–1444.

Choudhary, R. R., Meena, G., & Mohbey, K. K. (2022). Speech emotion based sentiment recognition using deep neural networks. *Journal of Physics, 2236*, 012003.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the twenty-second international joint conference on artificial intelligence (IJCAI)*. AAAI Press.

Daneshfar, F., Kabudian, S. J., & Neekabadi, A. (2020). Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier. *Applied Acoustics, 166*, 107360.

Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS)-younger talker_happy.

Dupuis, K., & Pichora-Fuller, M. K. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto emotional speech set. *Canadian Acoustics, 39*(3), 182–183.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Gomathy, M. (2021). Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm. *International Journal of Speech Technology, 24*(1), 155–163.

Gong, Y., Chung, Y.-A., & Glass, J. (2021). Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.

Gumelar, A. B., Yuniarno, E. M., Adi, D. P., Setiawan, R., Sugiarto, I., & Purnomo, M. H. (2022). Transformer-CNN automatic hyperparameter tuning for speech emotion recognition. In *2022 IEEE international conference on imaging systems and techniques (IST)* (pp. 1–6). IEEE.

Guo, L., Wang, L., Dang, J., Liu, Z., & Guan, H. (2019). Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. *IEEE Access, 7*, 75798–75809.

Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech*.

Han, S., Leng, F., & Jin, Z. (2021). Speech emotion recognition with a resnet-CNN-transformer parallel neural network. In *2021 International conference on communications, information system and computer engineering (CISCE)* (pp. 803–807). IEEE.

Huang, A., & Bao, P. (2019). Human vocal sentiment analysis. arXiv preprint arXiv:1905.08632.

Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 801–804).

Huang, Z.-W., Xue, W.-T., & Mao, Q.-R. (2015). Speech emotion recognition with unsupervised feature learning. *Frontiers of Information Technology & Electronic Engineering, 16*(5), 358–366.

Ismail, A., Idris, M. Y. I., Noor, N. M., Razak, Z., & Yusoff, Z. M. (2014). MFCC-VQ approach for qalqalahtajweed rule checking. *Malaysian Journal of Computer Science, 27*(4), 275–293.

Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control, 59*, 101894.

Jalal, M. A., Loweimi, E., Moore, R. K., & Hain, T. (2019). Learning temporal clusters using capsule routing for speech emotion recognition. In *Proceedings of interspeech 2019* (pp. 1701–1705). ISCA.

Jason, C. A., Kumar, S., et al. (2020). An appraisal on speech and emotion recognition technologies based on machine learning. *Language, 67*, 68.

Jiang, P., Fu, H., Tao, H., Lei, P., & Zhao, L. (2019). Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access, 7*, 90368–90377.

Karim, F., Majumdar, S., & Darabi, H. (2019). Insights into LSTM fully convolutional networks for time series classification. *IEEE Access, 7*, 67718–67725.

Kumaran, U., Radha Rammohan, S., Nagarajan, S. M., & Prathik, A. (2021). Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep c-rnn. *International Journal of Speech Technology, 24*(2), 303–314.

Kwon, S. (2019). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors, 20*(1), 183.

Kwon, S. (2021). Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *International Journal of Intelligent Systems, 36*(9), 5116–5135.

Kwon, S. (2021). MLT-Dnet: Speech emotion recognition using 1d dilated CNN based on multi-learning trick approach. *Expert Systems with Applications, 167*, 114177.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lee, J., & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Interspeech*.

Li, Y., Zhao, T., & Kawahara, T. (2019). Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech* (pp. 2803–2807).

Liu, M. (2022). English speech emotion recognition method based on speech recognition. *International Journal of Speech Technology, 25*(2), 391–398.

Livingstone, S. R., & Russo, F. A. (2018). 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, 13*(5), 0196391.

Li, S., Xing, X., Fan, W., Cai, B., Fordson, P., & Xu, X. (2021). Spatiotemporal and frequential cascaded attention networks for speech emotion recognition. *Neurocomputing, 448*, 238–248.

Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *29*.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia, 16*(8), 2203–2213.

Mao, Q., Xu, G., Xue, W., Gou, J., & Zhan, Y. (2017). Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. *Speech Communication, 93*, 1–10.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in science conference* (Vol. 8, pp. 18–25). Citeseer.

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D Log-Mel spectrograms with deep learning network. *IEEE Access, 7*, 125868–125881.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2227–2231). IEEE.

Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication, 49*(2), 98–112.

Mustaqeem, K. S. (2021). 1d-CNN: Speech emotion recognition system using a stacked network with dilated CNN features. *CMC-Computers Materials & Continua, 67*(3), 4039–4059.

Naqvi, R. A., Arsalan, M., Rehman, A., Rehman, A. U., Loh, W.-K., & Paul, A. (2020). Deep learning-based drivers emotion classification system in time series data for remote applications. *Remote Sensing, 12*(3), 587.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*, 8026–8037.

Praseetha, V., & Vadivel, S. (2018). Deep learning models for speech emotion recognition. *Journal of Computer Science, 14*(11), 1577–1587.

Rahaman, M. E., Alam, S. S., Mondal, H. S., Muntaseer, A.S., Mandal, R., & Raihan, M. (2019). Performance analysis of isolated speech recognition technique using MFCC and cross-correlation. In *2019 10th international conference on computing, communication and networking technologies (ICCCNT)* (pp. 1–4). IEEE.

Sajjad, M., Kwon, S., et al. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access, 8*, 79861–79875.

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech* (pp. 1089–1093).

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing, 1*(2), 119–131.

Singh, Y. B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. *Neurocomputing*.

Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., & Schuller, B. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5688–5691). IEEE

Tao, F., & Liu, G. (2018). Advanced LSTM: A study about better time dependency modeling in emotion recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2906–2910). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 30.

Wang, H., Zhang, Q., Wu, J., Pan, S., & Chen, Y. (2019). Time series feature learning with labeled and unlabeled data. *Pattern Recognition, 89*, 55–66.

Xu, X., Deng, J., Cummins, N., Zhang, Z., Wu, C., Zhao, L., & Schuller, B. (2017). A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(7), 1436–1449.

Zamil, A. A. A., Hasan, S., Baki, S. M. J., Adam, J. M., & Zaman, I. (2019). Emotion detection from speech signals using voting mechanism on classified frames. In *2019 International conference on robotics, electrical and signal processing techniques (ICREST)* (pp. 281–285). IEEE.

Zayene, B., Jlassi, C., & Arous, N. (2020). 3D convolutional recurrent global neural network for speech emotion recognition. In *2020 5th International conference on advanced technologies for signal and image processing (ATSIP)* (pp. 1–5). IEEE.

Zeng, M., & Xiao, N. (2019). Effective combination of densenet and BiLSTM for keyword spotting. *IEEE Access, 7*, 10767–10775.

Zeng, Y., Mao, H., Peng, D., & Yi, Z. (2019). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications, 78*(3), 3705–3722.

Zhang, S., Zhang, S., Huang, T., & Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia, 20*(6), 1576–1590.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1d & 2d CNN LSTM networks. *Biomedical Signal Processing and Control, 47*, 312–323.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.