International Journal of Speech Technology (2024) 27:319–327
https://doi.org/10.1007/s10772-024-10107-7

Feature fusion: research on emotion recognition in English speech

Yongyan Yang[1]

Received: 15 January 2024 / Accepted: 9 May 2024 / Published online: 30 May 2024

Abstract
English speech incorporates numerous features associated with the speaker's emotions, offering valuable cues for emotion recognition. This paper begins by briefly outlining preprocessing approaches for English speech signals. Subsequently, the Mel-frequency cepstral coefficient (MFCC), energy, and short-time zero-crossing rate were chosen as features, and their statistical properties were computed. The resulting 250-dimensional feature fusion was employed as input. A novel approach that combined gated recurrent unit (GRU) and a convolutional neural network (CNN) was designed for emotion recognition. The bidirectional GRU (BiGRU) method was enhanced through jump-joining to create a CNN-Skip-BiGRU model as an emotion recognition method for English speech. Experimental evaluations were conducted using the IEMOCAP dataset. The findings indicated that the fusion features exhibited superior performance in emotion recognition, achieving an unweighted accuracy rate of 70.31% and a weighted accuracy rate of 70.88%. In contrast to models like CNN-long short-term memory (LSTM), the CNN-Skip-BiGRU model demonstrated enhanced discriminative capabilities for different emotions. Moreover, it stood favorably against several existing emotion recognition methods. These results underscore the efficacy of the improved method in English speech emotion identification, suggesting its potential practical applications.

Keywords Feature fusion · English speech · Emotion recognition · Gated recurrent unit

1 Introduction

Beyond conveying textual information, the speaker's speech inherently carries emotional nuances such as joy and sadness. Even if the speaker articulates the same text, varying emotional cues can significantly alter the intended meaning. Hence, recognizing the emotional aspects of a speaker's speech is paramount (Liu et al., 2022). One of the differences between humans and computers is that humans have a good perception ability for emotions. The purpose of emotion recognition is to enable computers to simulate the process of human emotional perception, and speech, as a direct way of expressing emotions, plays an essential role in achieving human-computer interaction. Speech emotion identification has significant application value in many scenarios, for example, detecting the emotional states of drivers to issue timely reminders in situations of hyperactivity or fatigue (Requardt et al., 2020). It also applies in education (Tanko et al., 2022), aiding teachers in assessing students' emotional states through their speech. Moreover, in the realm of medicine and healthcare, speech-emotion recognition can be employed to discern if a patient is experiencing depression or anxiety (Hansen et al., 2021). For achieving intelligent and natural human-computer interaction (Pandey et al., 2022), extensive research has been conducted on emotion recognition in speech across different languages (Hu et al., 2021). Chattopadhyay et al. (2023) employed linear prediction coding and linear predictive cepstral coefficient extracted from speech signals as features and utilized clustering-based equilibrium optimizer and atom search optimization method for emotion recognition. They found that the method exhibited high classification accuracy. Guo et al. (2022) introduced a dynamic relative phase method for feature extraction. They employed a single-channel model and an attention-combined multi-channel model to learn acoustic features, yielding favorable results in emotion recognition experiments. Qiao et al.

✉ Yongyan Yang
yongyany@outlook.com

[1] Department of General Foreign Languages Education, Haikou University of Economics, Haikou, Hainan 571123, China

(2022) designed a Trumpet-6 method for the identification of emotions in Chinese speech, achieving a 95.7% accuracy in experiments on CASIA. Ocquaye et al. (2021) utilized a triple attentive CNN with asymmetric architecture for identifying emotion in cross-language speech. Experiments on English, German, and Italian datasets demonstrated the method's higher prediction accuracy. Given the widespread use of English (Hyder, 2021), research on emotion recognition in English speech holds significant practical value across various domains. The combination of CNN and long short-term memory (LSTM) or gated recurrent unit (GRU) has found extensive application in speech emotion recognition, such as CNN + LSTM (Ahmed et al., 2023), CNN-bidirectional gated recurrent unit (BiGRU) (Hu et al., 2022), and CNN-n-GRU (Nfissi et al., 2022), but there is still potential for further improvement in its performance. Building upon the combination of CNN and GRU, this paper proposes a recognition method to enhance its performance in English speech emotion recognition through feature fusion and structural improvements. By fusing features like energy and Mel-frequency cepstral coefficient (MFCC), richer emotional information was obtained. Furthermore, by incorporating skip connections, a Skip-BiGRU model was designed to combine with CNN, resulting in the CNN-Skip-BiGRU method for English speech emotion recognition. Its effectiveness was validated through experiments on IEMOCAP, providing a novel approach to differentiate emotions in English speech. This article provides some directions for further research on the integration of CNN with LSTM or GRU in speech emotion recognition and demonstrates the importance of feature fusion, providing some references for extracting speech emotion features.

2 Feature fusion in english speech

2.1 Preprocessing of speech signals

English speech signals must first be preprocessed to provide higher-quality speech for subsequent recognition. First, pre-emphasis is performed on original speech $x(n)$ by a first-order digital filter to make the spectrum flatter. The formula is written as:

$$y(n) = x(n) - \mu x(n-1) \tag{1}$$

where μ is the pre-emphasis factor, generally 0.97.

Based on the short-time smoothness characteristic of the signal, it is also necessary to intercept the original signal into shorter signals by framing, generally using the method of overlapping framing. After the framing, the key waveforms are highlighted by adding windows frame by frame, and in this paper, the Hamming window is used (Tan et al., 2020):

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left[2\pi n/(N-1)\right], 0 \le n \le N-1 \\ 0, other \end{cases} \tag{2}$$

The signal after adding the window is:

$$y(n) = \sum_{n=-N/2+1}^{N/2} x(m)\, w(n-m) \tag{3}$$

where n is the moment and N is the frame length.

2.2 Emotion feature extraction

Since the emotional information contained in a single feature tends to be one-sided, the following features are used for fusion to characterize the emotional information contained in the signal more comprehensively.

2.2.1 Energy

In general, the speaker's voice is louder when he is happy and angry and lower when he is sad and calm. The level of energy can reflect this difference. By utilizing the short-term average amplitude, it is possible to derive the energy features of the signal. The corresponding formula is:

$$E_n = \sum_{m=0}^{N-1} |x_n(m)| \tag{4}$$

where N is the frame length and $\mathrm{x_n(m)}$ means the n-th frame signal.

2.2.2 Short-time zero-crossing rate

It denotes the frequency at which the signal waveform crosses the zero level (Zhu et al., 2021). The number of times the signal passes the zero level varies depending on the emotional information contained in the signal. The formula is:

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1} |sgn\left[x_n(m)\right] - sgn\left[x_n(m-1)\right]| \tag{5}$$

$$sgn[x] = \begin{cases} 1, x \ge 0 \\ -1, x < 0 \end{cases} \tag{6}$$

2.2.3 Mel-frequency cepstral coefficient

MFCC is widely employed as a prevalent acoustic characteristic for examining auditory attributes of the human ear (Wibawa & Darmawan, 2021). It can help distinguish different emotional information. The correlation between Mel frequency and the true frequency is:

$$\mathrm{Mel}\,(\mathrm{f}) = 2595\mathrm{lg}\,(1 + \mathrm{f}/700) \tag{7}$$

The extraction process of MFCC is as follows.

① Fast Fourier transform (FFT) is performed on the signal:

$$X_j\,(k) = \sum_{i=0}^{N-1} x_j\,(n)\, e^{\frac{j2\pi nk}{N}}, 0 \le k \le K$$

② The signal passes through a set of Mel filters:

$$h_i\,(k) = \begin{cases} 0, k < f\,(i-1) \\ \frac{k-f(i-1)}{f(i)-f(i-1)}, f\,(i-1) \le k \le f\,(i) \\ \frac{f(i+1)-k}{f(i+1)-f(i)}, f\,(i) < k < f\,(i+1) \\ 0, k > f\,(i+1) \end{cases}$$

③ The logarithmic energy of the output of the Mel filter is calculated:

$$m\,(i) = \sum_{k=0}^{n-1} |X\,(k)|^2 h_i\,(k)\,, 0 \le i \le M$$

④ The logarithmic is taken from the outputs of all filters, and a discrete cosine transform (DCT) is also conducted to obtain the MFCC:

$$MFCC\,(i) = \sqrt{\frac{2}{M}} \sum_{i=0}^{M-1} lgm\,(i) \cos\left[(i - 1/2)\,\frac{i\pi}{M}\right]$$

In the above equations, $x_j\,(n)$ refers to the j-th frame of the English speech signal, K is the length of the FFT, which is 512, M is the number of filters, 24 filters in this paper, and $f\,(i)$ is the center frequency of the i-th filter.

2.2.4 Statistical characteristic

To obtain the emotional characteristics of the signal globally, this paper calculates the statistical features, including:

① mean value: $f_{mean} = \frac{1}{n}\sum_{i=1}^{n} f_i$;
② variance: $f_{var} = \frac{1}{n}\sum_{i=1}^{n} (f_i - f_{mean})^2$;
③ maximum value: $f_{max} = max\,(f_1, f_2, \cdots, f_n)$;
④ minimum value: $f_{min} = min\,(f_1, f_2, \cdots, f_n)$;
⑤ median: $f_{median} = \frac{f_{max}+f_{min}}{2}$

In the subsequent emotion recognition process, this paper selects the following features: energy, short-time zero-crossing rate, 24-dimensional MFCC, and 24-dimensional first-order difference dynamic feature ΔMFCC. These features are fused, resulting in a total of 50 dimensions. Subsequently, five statistical features are computed for the 50-dimensional feature, ultimately yielding a 250-dimensional feature.

3 Emotion recognition methods based on feature fusion

3.1 Convolutional neural network

CNNs are widely used to recognize images, text, speech, etc. (Ponmalar & Dhanakoti, 2022). In this paper, CNN is used to realize the learning of the 50-dimensional fused feature obtained in the previous section to get more advanced features. CNN has three main layer structures. Its structure is shown in Fig. 1.

(1) Convolutional layer: it is capable of autonomous learning of input English speech features. For an input feature matrix called I, if there is a $m \times n$ convolution kernel K, the convolution operation can be written as:

$$O_{i,j} = f\left(\sum_m \sum_n I_{i+m,j+n} K_{m,n} + w_b\right) \tag{8}$$

where $I_{i+m,j+n}$ means the element at the $(\mathrm{i}+\mathrm{m}, \mathrm{j}+\mathrm{n})$ of I, $K_{m,n}$ means the element at the (m, n) of K, and w_b is the bias.

(2) Pooling layer: The convolutional layer's output can be downsampled by this layer to capture the most salient feature (Li et al., 2019). Pooling operations can be divided into two types.

① Maximum pooling: Select the highest value from the local area as the output to obtain the most significant features.

② Mean pooling: Take the highest value in the local area as the output to obtain an average representation of the overall features.

(3) Fully connected layer: it synthesizes the features extracted from the first two layers to achieve recognition and classification.

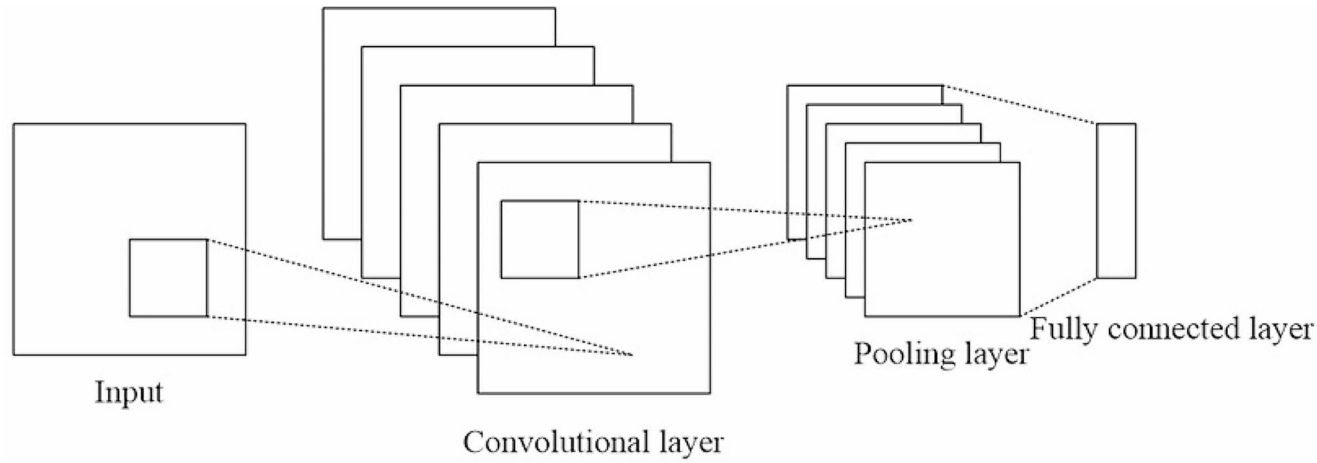
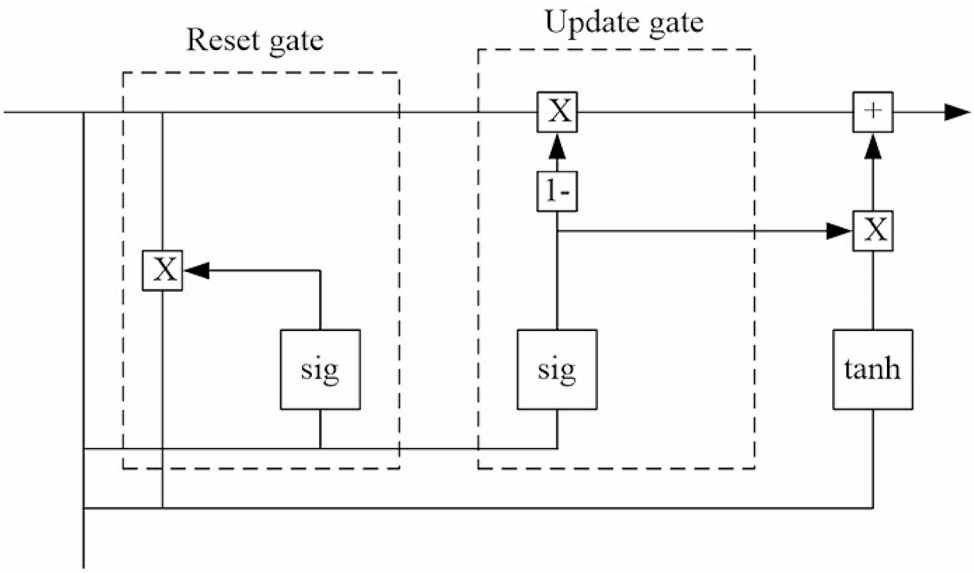


Fig. 1 The structure of CNN

Fig. 2 The structure of GRU



3.2 Bidirectional gated recurrent unit (BiGRU)

CNN can obtain more emotional features from the fused feature, but it is insufficient in the extraction of temporal context information; therefore, in this paper, the BiGRU model is used to learn temporal context information in English speech signals based on CNN. The BiGRU model utilizes both forward and backward GRU, enabling concurrent processing of past and future information (Niu et al., 2022). GRU exhibits a more streamlined architecture and superior training efficacy when compared to the long short-term memory network (LSTM) (Chen et al., 2021), and its structure is presented in Fig. 2.

According to Fig. 2, the update process of the reset gate can be written as:

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{9}$$

The update process of the update gate can be written as:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{10}$$

The output of GRU can be written as:

$$\tilde{h}_t = \tanh\left[W_h x_t + U_h\left(r_t \odot h_{t-1}\right) + b_h\right] \tag{11}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t \tag{12}$$

where x_t is the input, h_{t-1} is the previously hidden state, W and U are the weight matrices, b is the bias, σ is the sigmoi function, $\tilde{h}_t$ is the candidate output state, and h_t is the final GRU output state.

The hidden layer outputs of the forward GRU and the backward GRU can be obtained at the t moment:

$$\overrightarrow{h_t} = GRU\left(x_t, \overrightarrow{h_{t-1}}\right) \tag{13}$$

$$\overleftarrow{h_t} = GRU\left(x_t, \overleftarrow{h_{t-1}}\right) \tag{14}$$

They are combined to obtain the output of BiGRU at the t moment:

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{15}$$

3.3 Emotion recognition method based on CNN-Skip-BiGRU

To improve the effectiveness of the BiGRU model on long-term word sequence learning, this paper improves the structure of the BiGRU model by combining skip connections. Finally, it obtains a CNN-Skip-BiGRU model as an English speech emotion recognition method. Its structure is as follows.

(1) Input layer: the fused 250-dimensional English speech feature.
(2) CNN layer: it contains two convolutional layers and two pooling layers, all with a specification of 1×2 and a step length of 1.
(3) Skip-BiGRU layer (Fig. 3): it contains three BiGRU layers and uses skip connections, and the output of each layer is:

$$O_1 = GRU_1(x_t) \tag{16}$$

$$O_2 = GRU_2(O_1) \tag{17}$$

$$O_3 = GRU_3(O_1 + O_2) \tag{18}$$

(1) Dense layer: the features obtained from the above learning undergo dimensional variation to achieve a size of 256×64.
(2) Flatten layer: it flattens the multi-dimensional feature vector into one dimension.
(3) Softmax layer: it realizes the recognition of different English speech emotions, and the final output can be written as:

$$Y = softmax(flatten(Dense(O_1 + O_2 + O_3))) \tag{19}$$

4 Results and analysis

4.1 Experimental setup

The experiment was conducted on the Ubuntu 16.04 operating system. Python 3.6 was used as the programming language. The Keras platform was utilized to implement the emotion recognition approach, and the experiment employed the five-fold cross-validation method. The Adam optimizer was used. The batch size was 64, the epoch number was 150, and the initial learning rate was established as 10^{-4}. The IEMOCAP dataset was used (Ayadi & Lachiri, 2022), an English corpus recorded by ten professional performers. The samples were collected at a frequency of 16 kHz. The dataset was approximately 12 h and encompassed various emotion types such as happiness and anger. Due to category imbalance in the dataset, four emotions were selected for the experiments, and their distributions are shown in Table 1.

The effectiveness of the emotion recognition method was assessed using the following pair of indicators.

(1) Unweighted accuracy rate (UAR): it refers to the accuracy of the entire test set:

$$UAR = \frac{N_{acc}}{N} \tag{20}$$

where N is the total quantity of samples and N_{acc} is the count of accurately recognized specimens.

(2) Weighted accuracy rate (WAR): it represents the mean recognition accuracy for each emotion:

$$WAR = \frac{1}{n_{class}} \sum_{i=1}^{n_{class}} \frac{N_i^{acc}}{N_i} \tag{21}$$

where n_{class} refers to the number of emotion categories, N_i^{acc} refers to the recognition accuracy of the i-th kind of emotion, and N_i refers to the total number of samples for the i-th kind of emotion.

4.2 Results analysis

First, the effects of different features on the effect of English speech emotion recognition were analyzed, and the findings can be observed in Table 2.

From Table 2, the CNN-Skip-BiGRU model achieved a UAR of 63.24% and a WAR of 63.36% on the IEMOCAP dataset when using only MFCC-related features as inputs. This indicated that the model was less effective in recognizing different emotion types under these conditions. When

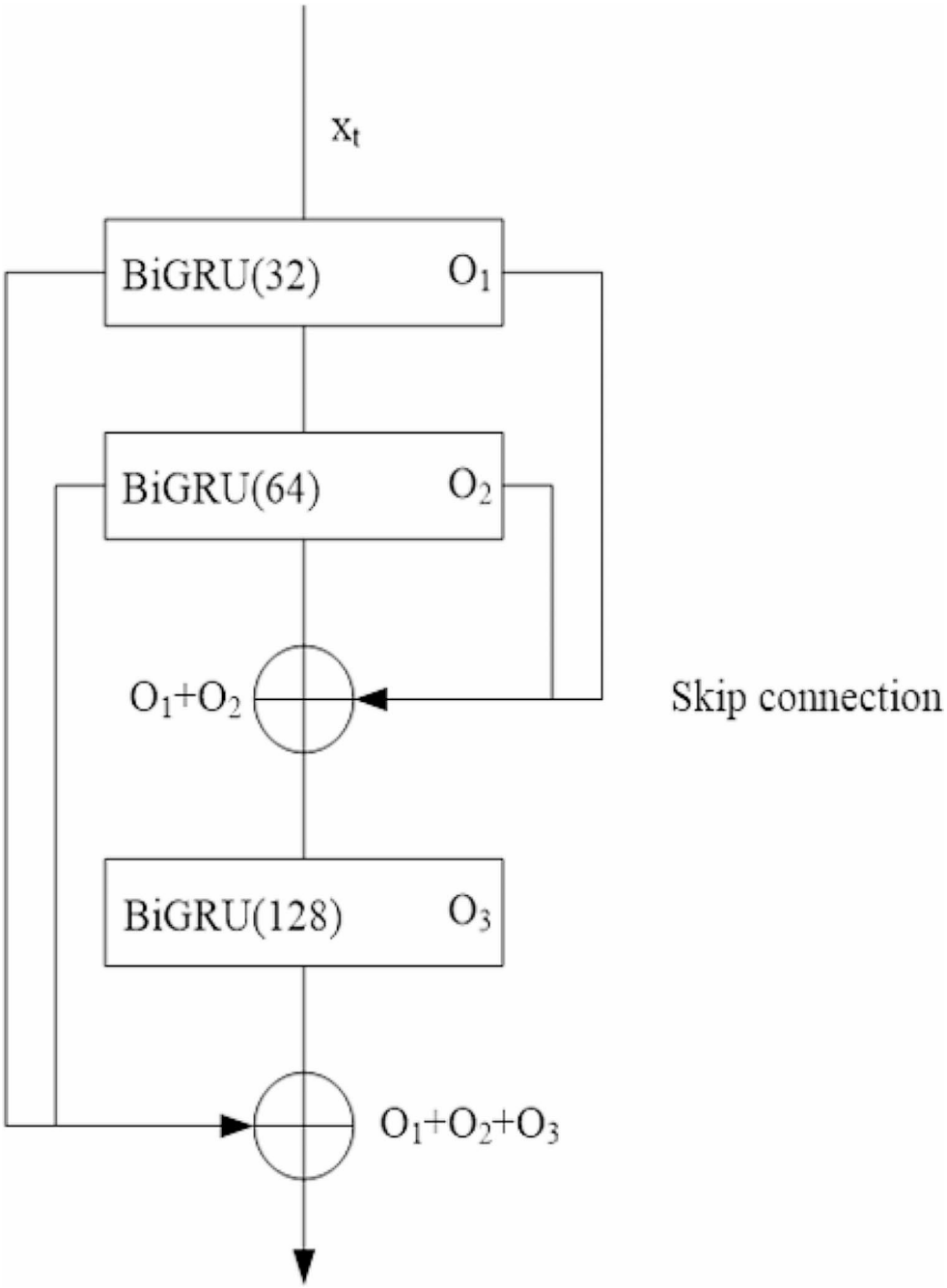


Fig. 3 Skip-BiGRU layer structure

Table 1 IEMOCAP dataset

Type of emotion	Sample size
Neutral	1,708
Happiness	1,636
Anger	1,103
Sorrow	1,084

Table 2 The impact of different features on the effect of English speech emotion recognition

	UAR/%	WAR/%
MFCC + △MFCC	63.24	63.36
MFCC + ΔMFCC + energy + short-time zero-crossing rate	67.45	66.97
The fused feature combining statistical features	70.31	70.88

fusing energy and short-time zero-crossing rate with MFCC-related features to obtain the 50-dimensional fused feature as input, the CNN-Skip-BiGRU model showed a UAR of 67.45% and a WAR of 66.97%, marking an improvement of

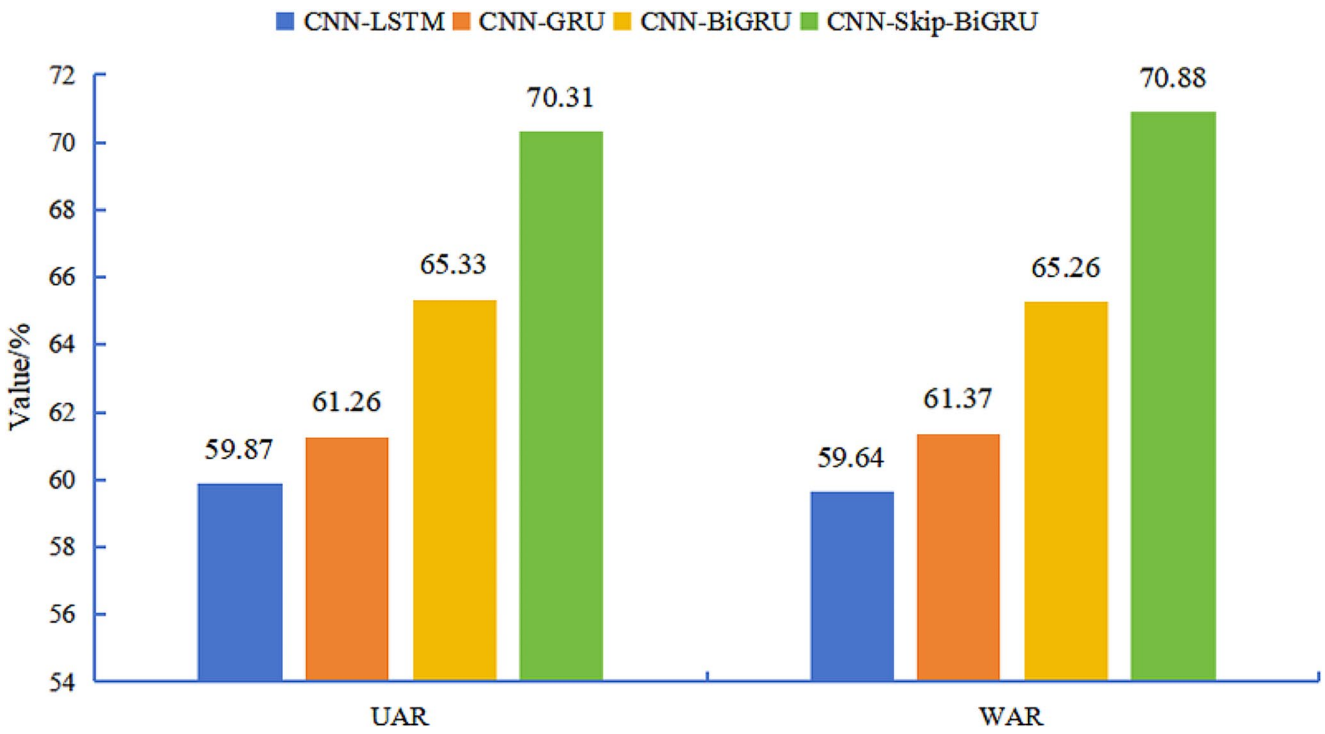


Fig. 4 Emotion recognition effect of different structural models

4.21% and 3.61% compared to using only MFCC, respectively. Finally, when using the obtained 250-dimensional feature as input, the UAR was 70.31%, and the WAR was 70.88%, showing improvements of 7.07% and 7.52%, respectively, compared to using the MFCC features alone. These results demonstrated the effectiveness of the fused features selected for English speech emotion identification.

The performance of the Skip-BiGRU structure was evaluated with the 250-dimensional feature as input (Fig. 4).

From Fig. 4, the combination of CNN and LSTM only achieved a UAR of 59.87% and a WAR of 59.64% on the IEMOCAP dataset, indicating that the model was weak in distinguishing between different emotion types. The UAR of CNN-GRU was 61.26%, and the WAR was 61.37%, which were improved by 1.39% and 1.73% respectively compared to the CNN-LSTM model, and this demonstrated the superiority of GRU over LSTM. Subsequently, the UAR of the CNN-BiGRU model was 65.33%, and the WAR was 65.26%, marking a further improvement of 4.07% and 3.89% compared to the CNN-GRU model. This result demonstrated the effectiveness of using BiGRU in feature learning. Finally, the CNN-Skip-BiGRU model attained a UAR of 70.31% and a WAR of 70.88%, surpassing the CNN-BiGRU model by 4.98% and 5.62%, respectively. This result indicated that using skip connections to optimize BiGRU significantly improved the effectiveness of the model in recognizing emotions in English speech.

Table 3 Comparisons with other methods

	UAR/%	WAR/%
3D-CRNNs	60.93	61.98
Attention-BLSTM-FCNs + DNN	60.10	59.70
ABLSTM-AFCN	67.00	68.10
CNN-Skip-BiGRU	70.31	70.88

The CNN-Skip-BiGRU model was compared with other emotion recognition methods:

(1) 3D-CRNNs (Peng et al., 2018): a 3D convolutional recurrent neural network-based method;
(2) attention-BLSTM-FCNs + DNN (Zhao et al., 2018): a method that combines the attention mechanism and bidirectional LSTM with a fully connected CNN to learn speech features and then utilizes a DNN to achieve sentiment prediction;
(3) ABLSTM-AFCN (Zhao et al., 2019): an approach that integrates an attention-combined bidirectional LSTM with an attention-combined fully convolutional network.

Refer to Table 3 for the comparative results.

From Table 3, it is observed that most current emotion recognition methods were based on deep learning, and they introduced more networks or the attention mechanism to the CNN-RN model to enhance the efficacy of emotion recognition. However, these attempts did not yield satisfactory

results. The attention-BLSTM-FCNs + DNN model was the least effective among the compared methods, achieving 60.10% UAR and 59.70% WAR, respectively. The ABLSTM-AFCN model performed relatively well with 67.00% UAR and 68.10% WAR, and the CNN-Skip-BiGRU model attained 70.31% UAR and 70.88% WAR, outperforming the other methods. This result indicated the reliability of the proposed method in English speech emotion identification, demonstrating its ability to distinguish between various emotions effectively.

5 Conclusion

This study proposes a CNN-Skip-BiGRU model for identifying emotions in English speech, which incorporates various features as inputs. Experiments on the IEMOCAP dataset revealed that the fused feature offered an effective characterization of emotion information in diverse English languages, leading to improved emotion recognition performance. Compared to LSTM and similar models, the Skip-BiGRU structure effectively enhanced the model's capability to distinguish between different emotions, outperforming other emotion recognition methods. These findings suggest that the designed CNN-Skip-BiGRU method holds promise for practical applications in real-world English speech emotion recognition.

However, this study also has some limitations. For instance, it only focuses on the recognition of four emotions in the IEMOCAP dataset and fails to further validate the practicality of the method on a wider range of languages and more diverse datasets. Therefore, future work should take into account the issue of dataset balance, verify the reliability of the proposed approach in recognizing a broader range of emotions, and conduct experiments on more extensive datasets.

Authors' contributions YYY conceived the idea for the study, did the analyses, and wrote the paper.

Funding Not applicable.

Data availability The data in this paper are available from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Ahmed, M. R., Islam, S., Islam, A. M., & Shatabda, S. (2023). An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. *Expert Systems with Applications*, *218*, 119633.

Ayadi, S., & Lachiri, Z. (2022). Visual emotion sensing using convolutional neural network. *Przeglad Elektrotechniczny*, *98*(3), 89–92.

Chattopadhyay, S., Dey, A., Singh, P. K., Ahmadian, A., & Sarkar, R. (2023). A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm. *Multimedia Tools and Applications*, *82*(7), 9693–9726.

Chen, Y., Liu, G., Huang, X., Chen, K., Hou, J., & Zhou, J. (2021). Development of a surrogate method of groundwater modeling using gated recurrent unit to improve the efficiency of parameter auto-calibration and global sensitivity analysis. *Journal of Hydrology*, *598*(3), 1–16.

Guo, L., Wang, L., Dang, J., Chng, E. S., & Nakagawa, S. (2022). Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition - ScienceDirect. *Speech Communication*, *136*, 118–127.

Hansen, L., Zhang, Y. P., Wolf, D., Sechidis, K., Ladegaard, N., & Fusaroli, R. (2021). A generalizable speech emotion recognition model reveals depression and remission. *Acta Psychiatrica Scandinavica*, *145*(2), 186–199.

Hu, D., Chen, C., Zhang, P., Li, J., Yan, Y., & Zhao, Q. (2021). A two-stage attention based modality fusion framework for multi-modal speech emotion recognition. *IEICE Transactions on Information and Systems*, *E104.D*(8), 1391–1394.

Hu, Z., Wang, L., Luo, Y., Xia, Y., & Xiao, H. (2022). Speech emotion recognition model based on attention CNN Bi-GRU fusing visual information. *Engineering Letters, 30*(2).

Hyder, H. (2021). The pedagogy of English language teaching using CBSE methodologies for schools. *Advances in Social Sciences Research Journal*, *8*, 188–193.

Li, Z., Wang, S. H., Fan, R. R., Cao, G., Zhang, Y. D., & Guo, T. (2019). Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling. *International Journal of Imaging Systems and Technology*, *29*(4), 577–583.

Liu, L. Y., Liu, W. Z., Zhou, J., Deng, H. Y., & Feng, L. (2022). ATDA: Attentional temporal dynamic activation for speech emotion recognition. *Knowledge-based Systems*, *243*(May 11), 1–11.

Nfissi, A., Bouachir, W., Bouguila, N., & Mishara, B. L. (2022). CNN-n-GRU: End-to-end speech emotion recognition from raw waveform signal using CNNs and gated recurrent unit networks. In *21st IEEE international conference on machine learning and applications (ICMLA)*, (pp. 699–702).

Niu, D., Yu, M., Sun, L., Gao, T., & Wang, K. (2022). Short-term multi-energy load forecasting for integrated energy systems based on CNN-BiGRU optimized by attention mechanism. *Applied Energy*, *313*, 1–17.

Ocquaye, E. N. N., Mao, Q., Xue, Y., & Song, H. (2021). Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *International Journal of Intelligent Systems*, *36*(1), 53–71.

Pandey, S. K., Shekhawat, H. S., & Prasanna, S. R. M. (2022). Attention gated tensor neural network architectures for speech emotion recognition. *Biomedical Signal Processing and Control*, *71*(2), 1–16.

Peng, Z., Zhu, Z., Unoki, M., Dang, J., Akagi, M. (2018). Auditory-inspired end-to-end speech emotion recognition using 3D convolutional recurrent neural networks based on spectral-temporal representation. In *2018 IEEE international conference on, multimedia, & expo. (ICME)* (pp. 1–6), San Diego, CA, USA.

Ponmalar, A., & Dhanakoti, V. (2022). Hybrid whale tabu algorithm optimized convolutional neural network architecture for intrusion detection in big data. *Concurrency and Computation: Practice and Experience*, *34*(19), 1–15.

Qiao, D., Chen, Z. J., Deng, L., & Tu, C. L. (2022). Method for Chinese speech emotion recognition based on improved speech-processing convolutional neural network. *Computer Engineering*, *48*(2), 281–290.

Requardt, A. F., Ihme, K., Wilbrink, M., & Wendemuth, A. (2020). Towards affect-aware vehicles for increasing safety and comfort: Recognising driver emotions from audio recordings in a realistic driving study. *IET Intelligent Transport Systems*, *14*(10), 1265–1277.

Tan, M., Wang, C., Yuan, H., Bai, J., & An, L. (2020). FDA-MIMO Beampattern synthesis with Hamming window weighted linear frequency increments. *International Journal of Aerospace Engineering*, *2020(2)*, 1–8.

Tanko, D., Dogan, S., Demir, F. B., Baygin, M., Sahin, S. E., & Tuncer, T. (2022). Shoelace pattern-based speech emotion recognition of the lecturers in distance education: ShoePat23. *Applied Acoustics*, *190*, 1–9.

Wibawa, I. D. G. Y. A., & Darmawan, I. D. M. B. A. (2021). Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini. *Journal of Physics: Conference Series, 1722*, 1–8.

Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., & Li, C. (2018). Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. In *Annual conference of the international speech communication association*, (pp. 272–276).

Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., & Schuller, B. (2019). Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for Speech emotion recognition. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 97515–97525.

Zhu, M., Cheng, J., & Zhang, Z. (2021). Quality control of microseismic P-phase arrival picks in coal mine based on machine learning. *Computers & Geosciences*, *156*, 1–12.