Citation: Liyan Zhang, Jiaxin Du, Shuang Chen, et al. Improved MFCC features and TWM model for speech emotion recognition. *Journal of Harbin Institute of Technology (New Series)*. DOI: 10.11916/j.issn.1005-9113.24051

Improved MFCC Features and TWM Model for Speech Emotion Recognition

Liyan Zhang[*1], *Jiaxin Du*[1,2], *Shuang Chen*[1] *and Jiayan Li*[1]

(*1.School of computer and communication engineering, Dalian Jiaotong University, Dalian, Liaoning, 116028, China; 2.School of Information Engineering, Zhengzhou University of Industrial Technology, Zhengzhou, Henan, 451100, China*

Abstract: To solve the problem that traditional MFCC features cannot fully represent dynamic speech features, this paper introduces first-order and second-order difference on the basis of static MFCC features to extract dynamic MFCC features, and constructs a hybrid model (TWM, TIM-NET WGAN-GP multi-head Attention) combining multi-head attention mechanism and improved Wasserstein Generative Adversarial Network (WGAN-GP) on the basis of TIM-NET network. Among them, the multi-head attention mechanism not only effectively prevents gradient vanishing, but also allows for the construction of deeper networks that can capture long-range dependencies and learn from information at different time steps, improving the accuracy of the model; WGAN-GP solves the problem of insufficient sample size by improving the quality of speech sample generation. The experimental results show that this method significantly improves the accuracy and robustness of speech emotion recognition on RAVDESS and EMO-DB datasets.

Keywords: dynamic features; speech emotion recognition; multi-head attention mechanism; generative adversarial networks

CLC number: TP183 **Document code**: A **Article ID**: 1005-9113(2025)00-0000-09

0 Introduction

With the continuous development of technology, human-computer interaction technology is becoming increasingly mature. However, current machines are still unable to effectively recognize human emotions, making speech emotion recognition a hot research topic[1]. The main problems faced by this technology include: 1) effectively extracting speech emotional features; 2) Building an accurate emotion recognition model; 3) Emotional voice data that requires diversity and complexity.

Early research on speech emotion recognition mainly relied on traditional acoustic features, which can be divided into prosodic features, spectral based features, and speech quality features[2]. Therefore, researchers began to explore more effective features. Liu et al.[3] improved the Gammatone Frequency Cepstral Coefficient (GFCC) and proposed the VGFCC feature. Kumaran et al.[4] combined MFCC (Mel Frequency Cepstral Coefficient) and GFCC features and used Convolutional Recurrent Neural Network (CRNN) for recognition. MFCC is a commonly used spectral based feature in speech emotion recognition. However, traditional MFCC features mainly reflect the static characteristics of speech signals and cannot effectively capture the dynamic changes of speech signals, which may affect the accuracy of emotion recognition. To address this issue, this paper introduces a novel approach that combines MFCC features with their first and second derivatives[5] to capture both the static and dynamic changes of speech signals, providing a more comprehensive feature representation and improving the performance of speech emotion recognition models.

In terms of constructing emotion recognition models, TIM-NET (temporal-aware bi-directional multi-scale network), as an advanced network architecture, has demonstrated good performance in various emotion recognition tasks. TIM-NET better

Received 2024-08-20.

* Corresponding author: Liyan Zhang, Ph.D candidate. Email: zhangliyan@126.com.

adapts to emotional changes by fusing features from multiple time scales. However, dealing with complex emotional data, TIM-NET still faces the problem of insufficient ability to process long time series. This is due to its time step limitation, which makes the modeling of long-term dependencies insufficient. To overcome this issue, this paper introduces an innovative solution by incorporating a multi-head attention mechanism into the TIM-NET architecture. The multi-head attention mechanism helps capture long-range dependencies and enables the construction of deeper networks, where each attention head focuses on different aspects of the data. This significantly enhances the networks ability to handle complex relationships within the data.

In response to the lack of diversity and complexity in sentiment data, this paper adopts an improved generative adversarial network WGAN-GP (Wasserstein GAN with Gradient Penalty) to generate more diverse and complex data samples. WGAN-GP solves the common problems of pattern collapse and training instability in traditional GANs by introducing gradient penalty mechanism. The gradient penalty mechanism limits the discriminator's ability to generate gradient changes in the samples, making the training process between the generator and discriminator more stable. This study is able to generate diverse emotional speech data, effectively improving the model's generalization ability and recognition accuracy.

This article constructs a hybrid model (TWM, TIM-NET WGAN-GP Multi-Head Attention) based on TIM-NET network, which combines multi-head attention mechanism and improved Wasserstein Generative Adversarial Network (WGAN-GP), and combines improved MFCC features for speech emotion recognition. This model represents significant advancements in feature extraction, emotion modeling, and data generation, effectively improving the robustness and accuracy of speech emotion recognition.

1 TIM-NET Network Model

Temporal-aware bi-directional multi-scale network (TIM-NET) is capable of learning long-term emotional dependencies from both forward and backward directions, and capturing multi-scale features at the frame level. The TIM-NET network architecture is shown in Fig. 1[6].

Firstly, using MFCC as the input of the network, each speech signal is divided into 50 ms frames and subjected to 2048 point Fourier transform through frame segmentation and Hamming window operation. Then, Mel scale triangular filter bank analysis is performed to extract the first 39 factors to obtain low-frequency features. TIM-NET consists of several forward and backward Temporal Aware Blocks (TAB), each TAB consisting of two sub blocks and a Sigmoid function σ. The Sigmoid function is used to learn the Temporal Attention Map A and generate the Temporal Aware Feature F by multiplying it element by element with the input. The sub-blocks of TAB use Dilated Causal Convolutions (DC Conv) with causal constraints to expand the acceptance range and avoid information loss, followed by batch normalization, ReLU (Rectified Linear Unit) activation, and spatial dropout operations to enhance the model's nonlinear features and generalization ability[7]. Node time modeling determines emotional categories by integrating complementary information from the past and future. The forward and backward processing processes are shown in Eqs. (1) and (2), respectively, combining bidirectional semantic dependency and discourse level global context representation, where $\overrightarrow{F_1}$ the output from the first 1×1 convolution layer, the global time pool operation.

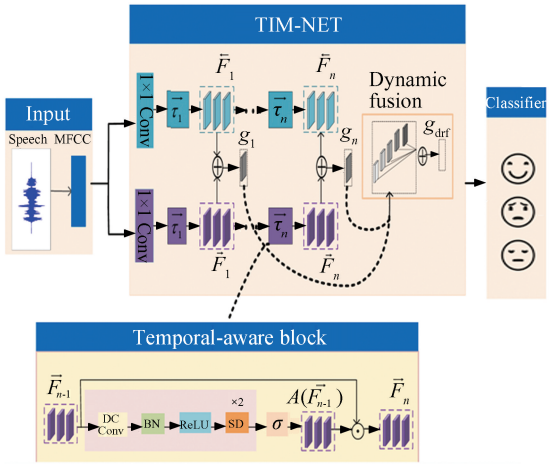

Fig. 1 TIM-NET network structure diagram

The entire process is shown in Eq. (3), where the global time pooling operation G takes the average of the time dimension and generates a representation vector of a specific receptive domain from the j^{th} TAB. In order to adapt to the temporal scale changes in

pronunciation habits, TIM-NET has designed a multi-scale dynamic module. During the training phase, the module selects the appropriate time scale based on the current input and fuses the features with the Dynamic Receptive Field (DRF) through weighted summation. The weight fusion w DRF comes from different TAB, and the DRF fusion g_{drf} is shown in Eq. (4), where $w_{\mathrm{drf}} = [w_1, w_2, \cdots, w_n]^T$ are trainable parameters. Once the emotion representation w_{drf} has strong discriminability, the fully connected layer of the softmax function can be utilized for emotion classification.

$$\overrightarrow{F_{j+1}} = A(\overrightarrow{F_j}) \odot \overrightarrow{F_j} \tag{1}$$

$$\overleftarrow{F_{j+1}} = A(\overleftarrow{F_j}) \odot \overleftarrow{F_j} \tag{2}$$

$$g_j = G(\overleftarrow{F_j} + \overleftarrow{F_j}) \tag{3}$$

$$g_{\mathrm{drf}} = \sum_{j=1}^{n} w_j g_j \tag{4}$$

here, forward feature representation ($\overrightarrow{F_j}$) and backward feature representation ($\overleftarrow{F_j}$) at step or layer n capturing temporal information in different directions. $A(\cdot)$ represents an activation or transformation operation that introduces non-linearity for complex pattern learning. j represents an index or step counter that marks the sequence of operations or layers (for example, the j-th TAB). g_j represents feature fusion result at step or index j, obtained by processing the sum of forward and backward features with G. G represents a function for feature processing or fusion, acting on the sum of forward and backward features. w_j represents a trainable weight parameter that weighs g_j' contribution to the final dynamic receptive field fusion result g_{drf}.

Although TIM-NET has processed the dynamic features to some extent of time series through time aware blocks and multi-scale dynamic modules, its feature fusion may not fully capture all changes in time steps when facing very long time series. This time step limitation may result in information loss or insufficiency when TIM-NET processes speech data with long time spans, thereby limiting its ability to model long-term dependencies. This means that when processing long time series, TIM-NET may not be able to effectively capture all relevant emotional information, affecting its ability to process complex emotional data.

2 TWM Network Model with Improved MFCC Features

The construction of the TWM model in this article is based on the TIM-NET network model, which integrates multi-head attention mechanism and an improved WGAN-GP. Furthermore, it integrates the first and second moments of MFCC features. The network structure diagram is shown in Fig. 2.

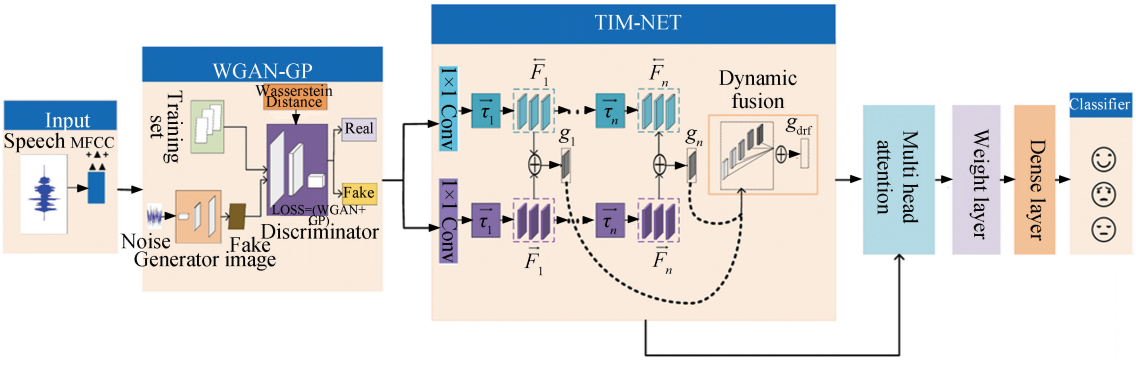

Fig. 2 TWM network model with improved MFCC features

Firstly, a generative adversarial network (WGAN-GP) is used for data augmentation. The generator is responsible for generating fake emotional data samples, while the discriminator is responsible for distinguishing between real samples and generated samples. WGAN-GP introduces Wasserstein distance to measure the difference between generated samples and real samples, and adds gradient penalty term in the loss function to improve the quality of generated samples and the stability of training. The generator and discriminator are alternately optimized during the training process, so that the generated samples gradually approach the distribution of the real samples. These generated fake samples will be used together with real samples to enhance the training dataset, in order to improve the robustness and

generalization ability of the TIM-NET model. Subsequently, the real data containing MFCC and its Delta and Delta Delta features, as well as the generated fake data, are input into the TIM-NET module. The TIM-NET module uses Dilated Causal Convolution (DC conv) with causal effects to extract temporal features. Extended convolution expands the receptive field of the convolution kernel in the time dimension by introducing larger time span intervals inside the kernel, thus better capturing long-term temporal dependencies. By combining these dilated convolutional layers, TIM-NET can extract multi-scale temporal features, thereby enhancing the ability of feature representation. Afterwards, the temporal information and multi-scale features extracted by TIM-NET are input into the multi-head attention mechanism. The multi-head attention mechanism processes different feature subspaces in parallel through multiple attention heads to capture the complex relationships between features. Each attention head weights and aggregates input features to generate multi-dimensional weighted feature representations, thereby enhancing the model's ability to model long-term dependencies and feature relationships, and effectively solving the problem of time step limitations[8]. After the output of the multi-head attention mechanism, a weight layer is added. This layer further processes features by matrix multiplying the output of the attention mechanism with a trainable weight matrix, and outputs subsequent feature representations. The output of the weight layer is passed to the fully connected layer (Dense), and the Softmax activation function is used for sentiment classification to obtain the final sentiment classification result.

2.1 Extraction of Improved MFCC Feature

Due to the human ear's greater sensitivity to the dynamic characteristics of sound signals, and since MFCC only reflects the static features of speech signals, the first-order and second-order derivatives are extracted after obtaining the MFCC features. The first-order derivative represents the relationship between the current speech frame and the previous frame, while the second-order derivative represents the relationship between the first-order and the second-order derivatives. Combining the static MFCCs with their first and second-order derivatives creates a more comprehensive representation of the speech signal, better capturing the characteristics of different emotional states.

This study improved the MFCC features by introducing the first derivative Δ (Delta) and second derivative ΔΔ (Delta Delta) features of MFCC to better capture the dynamic changes of speech signals. Specifically, the Δ (Delta) feature captures the dynamic information of the speech signal by calculating the time first derivative of the MFCC feature, as shown in Eq. (5), while the Δ (Delta Delta) feature describes the gradient changes of the speech signal by calculating the time second derivative of the MFCC feature, as shown in Eq. (6).

$$\Delta_t = \frac{\sum_{n=-N}^{N} n \cdot X_{t+n}}{\sum_{n=-N}^{N} n^2} \tag{5}$$

$$\Delta\Delta_t = \frac{\sum_{n=-N}^{N} n^2 \cdot \Delta_{t+n}}{\sum_{n=-N}^{N} n^4} \tag{6}$$

here, Δ_t represents the Δ feature of the t-th frame, Δ_t represents the $\Delta\Delta$ feature of the t-th frame, $\Delta\Delta_{t+n}$ is the MFCC feature of time frame $t + n$, and N is the window size of the ΔΔ feature (usually 2 or 3 frames). The original MFCC features, Δ features, and Δ features obtained through calculation are concatenated in the feature dimension to form a feature matrix containing multiple dynamic information. This method enhances the expression ability and accuracy of features, thereby improving the performance of speech emotion recognition.

2.2 WGAN-GP

Traditional GAN training often faces challenges of instability and mode collapse, mainly because it uses Jensen Shannon divergence as the loss function to measure the difference between the generated sample distribution and the true sample distribution. However, this measure may result in weak training signals and unstable training when the generated sample distribution differs significantly from the true sample distribution. WGAN addresses these issues by introducing Wasserstein distance as an optimization objective. The Wasserstein distance provides a more stable metric for calculating the difference between the generated sample distribution and the true sample distribution, as it directly measures the minimum moving cost between the two distributions. This measurement method can provide meaningful gradients, even though the sample distribution differences are small, which makes the training process more stable. However, WGAN still faces

some challenges, especially unstable discriminator training. WGAN requires the discriminator's function to satisfy 1-Lipschitz continuity, which means that the discriminator's gradient must remain within a certain range across all data points[9]. If the gradient of the discriminator is not controlled, it may lead to instability and non-convergence during the training process. WGAN-GP (WGAN with Gradient Penalty) further improves WGAN by introducing gradient penalty terms to enhance training stability. The gradient penalty term aims to ensure that the gradient of the discriminator maintains 1-Lipschitz continuity on all data points, thereby preventing unstable training of the discriminator. The calculation method of this gradient penalty term is shown in Eq. (7).

$$L_{\mathrm{GP}} = \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2] \quad (7)$$

here, $\mathbb{E}_{\hat{x}\sim P_{\hat{x}}}$ represents the expected calculation of the sample $\hat{x}$ extracted from the distribution $P_{\hat{x}}$. $\hat{x}$ is a sample interpolated between real samples and generated samples, i.e. $\hat{x} = \in x_{real} + (1 - \in) x_{fake}$, where x_{real} is the real sample and x_{fake} is a randomly generated sample within the range of $[0,1]$. $\nabla_{\hat{x}} D(\hat{x})$ represents the gradient of discriminator D for sample $\hat{x}$ $\| \cdot \|_2$ is the L2 norm, which is the magnitude of the gradient. A gradient penalty term $(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2$ is used to ensure that the magnitude of the gradient is close to 1, thereby maintaining Lipschitz continuity.

2.3 Multi-Head Attention Mechanism

The multi-head attention mechanism processes different feature subspaces in parallel through multiple attention heads to capture the complex relationships between features. Each attention head calculates its own weighted sum, and then concatenates these weighted sums to generate the final output[8]. This mechanism can better establish the interaction and long-term dependence between model features. The calculation process of the multi-head attention mechanism first performs linear transformation on the input query vectors $\boldsymbol{Q}, \boldsymbol{K}$, and $\boldsymbol{V}$, and calculates the scaled dot product attention of each head, as shown in Eq. (8).

$$\mathrm{Attention}(Q', K', V') = \mathrm{softmax}\left(\frac{Q' K' T}{\sqrt{d_k}}\right) V' \quad (8)$$

here, d_k represents the dimension of the key vector. Apply scaling to ensure that the input to the softmax function remains within a stable range, especially when the dimensionality of the key vectors is large.In order to learn the relevant information of emotional features from different subspaces, this paper uses multiple attention heads, each of which performs independent linear transformations on $\boldsymbol{Q}, \boldsymbol{K}$, and $\boldsymbol{V}$, to obtain multiple attention outputs. The output Q_{i} of the *i-th* head is calculated as shown in Eq. (9).

$$Q_i = \mathrm{Attention}(Q\, W_Q^i, K\, W_K^i, V\, W_V^i) \quad (9)$$

here, $W_i^Q \in R^{d_{\mathrm{model}}\times d_q}$, $W_i^K \in R^{d_{\mathrm{model}}\times d_k}$, $W_i^V \in R^{d_{\mathrm{model}}\times d_v}$. $R^{d_{\mathrm{model}}\times d_q}$, $R^{d_{\mathrm{model}}\times d_k}$, $R^{d_{\mathrm{model}}\times d_v}$ denote real-valued matrices, where d_{model} is the dimension of the model's feature space, and d_q、d_k、d_v are the dimensions of the query, key, and value spaces respectively. These matrices define the space for linear transformations in multi-head attention. W_i^Q、W_i^K、W_i^V are trainable weight matrices for the *i*-th attention head. W_i^Q transforms the input Q into the query space of the *i*-th head, W_i^K transforms K into the key space, and W_i^V transforms V into the value space. These weights are learned during training to capture different aspects of feature relevance in multi-head attention. Finally, concatenate the attention outputs of each head to obtain the final output of multi-head attention (MHA), as shown in Eq. (10).

$$MHAt(Q,K,V) = \mathrm{Concatenate}(Q_1, Q_2, \cdots, Q_N) \quad (10)$$

The multi-head attention mechanism can capture different features of input sequences from multiple subspaces, enhancing the network's ability to model long-range dependencies. Introducing multi-head attention mechanism in TIM-NET network can effectively alleviate the time step limitation problem faced by the original network when processing long-term emotional data, thereby improving the overall performance of the model[10].

3 Experiment

3.1 Data Set

The experiment used two datasets, RAVDESS and EMO-DB. The RAVDESS dataset contains 1440 speech files, covering 8 emotion categories including anger, neutral, calm, happy, sad, fearful, disgust, and surprise[11]. The EMO-DB dataset is a German emotional speech database recorded by the Technical University of Berlin, containing 535 voices. The emotional categories include neutral, anger, fear, happiness, sadness, disgust, and boredom[12].

3.2 Parameter Setting and Evaluation Indicators

In this study, the model was built using TensorFlow and Keras frameworks. The optimizer

selected Adam, the activation function was ReLU, the dropout rate was set to 0.1, the iteration cycle was 200 times, and a five-fold cross validation was performed. Finally, the performance of the model was represented by the mean of five results, and the batch size was set to 64. In the experiment, the most commonly used evaluation metrics in the field of Speech Emotion Recognition (SER) were used: weighted accuracy (WAR) and unweighted accuracy (UAR).

3.3 Experimental Results and Analysis

3.3.1 *Experimental comparison of improving MFCC features*

In order to verify the effectiveness of improving MFCC features, this study constructed the following three feature sets and combined them with the TWM model to conduct ablation experiments on RAVDESS and EMO-DB corpora[13]. The specific feature set is described as follows:

S_1: Only includes MFCC features. S_2: Composed of MFCC and its first-order differential features. S_3: Composed of MFCC and its first-order differential and second-order differential features.

These feature sets are used to analyze the performance of TWM models in speech emotion recognition tasks with different feature combinations. The experimental results are shown in Table 1, and evaluation measures are UAR (%) / WAR (%).

Table 1 Performance of different feature sets on RAVDESS and EMO-DB corpus

Features	Dimension	RAVDESS(%)	EMO-DB(%)
S_1	39	89.54/89.06	91.94/90.81
S_2	78	90.13/89.88	92.11/91.20
S_3	117	93.00/93.15	93.92/93.60

From the experimental results, it can be seen that as the dimensionality of the feature set increases, both the unweighted accuracy (UAR) and weighted accuracy (WAR) of the model significantly improve. This indicates that by introducing first-order differential (Delta) and second-order differential (Delta Delta) features, dynamic information in speech signals can be better captured, thereby improving the accuracy of speech emotion recognition. Among the three feature sets, the S_3 feature set performed the best, achieving 93.00% UAR and 93.15% WAR on the RAVDESS corpus, while achieving 93.92% UAR and 93.60% WAR on the EMO-DB corpus, respectively. This indicates that combining MFCC and its first-order and second-order differential features in the TWM model can obtain a more comprehensive representation of speech signal features, thereby improving the performance of the model.

3.3.2 *Ablation experiment*

To validate the effectiveness of the TWM model proposed in this study, improved MFCC features were used as inputs, and EMO-DB and RAVDESS datasets were used for training and evaluation. Weighted accuracy (WAR) and unweighted accuracy (UAR) were used as metrics, and TIM-NET was used as the baseline model to verify the impact of different modules on model performance. The experimental results are shown in Table 2, evaluation measures are UAR (%) / WAR (%).

From Table 2, it can be seen that the TWM model, which has a parameter count of 13 M, performs better than TIM-NET (2M), TIM-NET + WGAN-GP (12M), and TIM-NET + MHA (3M) on both RAVDESS and EMO-DB datasets. In both datasets, the UAR and WAR of the TWM model reached their highest values. This result validates the effectiveness of the TWM model in integrating the Multi-Head Attention mechanism (MHA) and the generative adversarial network (WGAN-GP). The TWM model delivers significant performance improvements in emotion recognition tasks over the TIM-NET baseline (2M-parameter), despite its increase in model size. The inclusion of WGAN-GP and MHA not only enhanced model performance but also increased the model's capacity for learning complex emotional features.

Table 2 Performance comparison of models on different datasets

Model	Parameters (M)	RAVDESS(%)	EMO-DB(%)
TIM-NET	2	89.96/89.61	91.79/90.21
TIM-NET+ WGAN-GP	12	90.04/89.97	92.20/91.71
TIM-NET+ MHA	3	90.88/88.13	91.95/92.26
TWM	13	93.00/93.15	93.92/93.60

3.3.3 *Model robustness verification*

To verify the robustness of the TWM model, Gaussian noise was introduced during the training and testing phases to evaluate the model's performance under data perturbations. The introduction of noise simulates real-world data interference, thereby

validating the model's robustness and reliability. This study used Gaussian noise with a standard deviation of 0.01 to perturb the input data. The chosen standard deviation ensures that the noise impact remains within a reasonable range, avoiding excessive interference with the original signal[14]. The validation was performed on the RAVDESS and EMO-DB datasets, and the corresponding confusion matrices for various emotions are shown in Figs. 3 and 4.

As observed in the confusion matrix in Fig. 3, without the addition of Gaussian noise, the TWM model achieves high classification accuracy across all categories. The model's performance is balanced and consistent, demonstrating its strong ability to accurately classify emotions. In contrast, Fig. 4 shows the confusion matrices with Gaussian noise, where the TWM model still maintains high classification accuracy despite the data perturbations. This proves the robustness of the TWM model, confirming its effectiveness in handling noisy data and its superior performance under real-world conditions.

3.3.4 Compared with existing methods

Zhao et al.[15] used Mel spectrograms and MFCC as audio description methods, and proposed a fully convolutional neural network architecture as a classifier. This method achieved an average accuracy of 75.28% on the RAVDESS dataset and 92.71% on the EMO-DB dataset. Jahangir et al.[16] used data augmentation and feature fusion methods, combined with a CNN model, and achieved accuracy of 90.60% and 93.30% on the RAVDESS dataset and EMO-DB dataset, respectively. Zhang et al.[11] proposed a speech emotion recognition method based on Dilated Residual Network (DRN) combined with auxiliary classifier and channel spatial attention fusion.

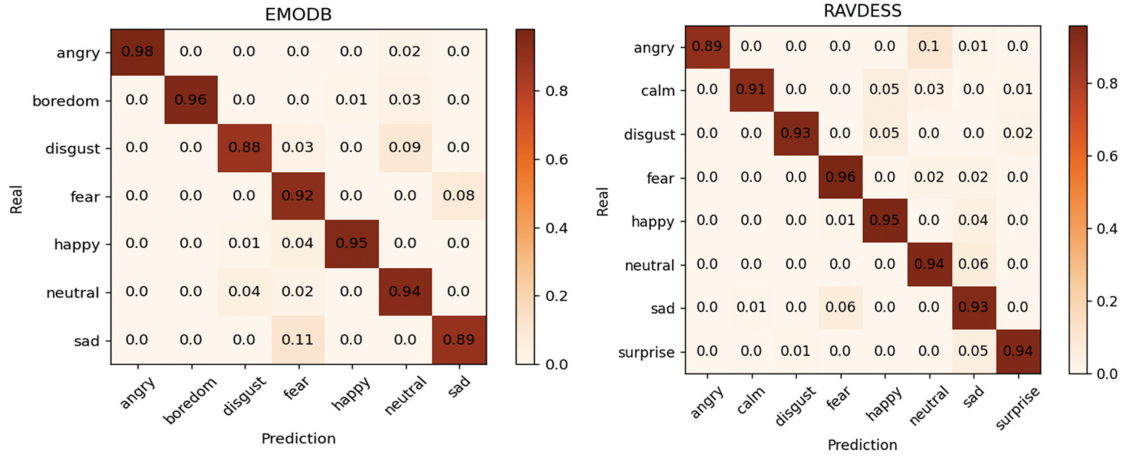

Fig. 3 Emotion classification confusion matrix of RAVDESS and EMO-DB datasets

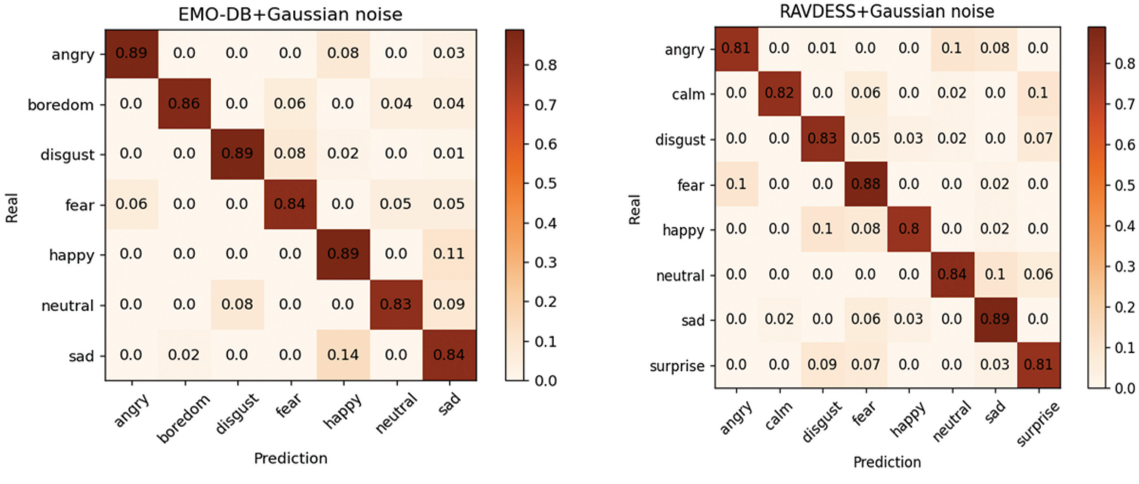

Fig. 4 Emotion classification confusion matrices of the RAVDESS and EMO-DB datasets with Gaussian noise

The model achieved an accuracy of 92.91% on the RAVDESS dataset and 89.15% on the EMO-DB dataset, demonstrating its effectiveness and generalization ability. García-Ordás et al.[17] used the Mel spectrogram and MFCC as audio description methods, and proposed a fully convolutional neural network architecture as a classifier, achieving an average accuracy rate of 75.28% on the RAVDESS dataset and an average accuracy rate of 92.71% on the EMODB dataset. Zhou et al.[18] utilized MFCC as the model input features and designed dual-path temporal convolutional channels based on Temporal Convolutional Networks (TCN) and cross-gated mechanisms to extract multi-scale cross-fusion features. The model achieved an average accuracy of 87.32% on the RAVDESS dataset and 89.30% on the EMODB dataset. In order to verify the effectiveness and robustness of the proposed method, this paper compared the proposed model with the advanced speech emotion recognition methods mentioned above, and conducted comparative analysis on the RAVDESS and EMO-DB datasets. As shown in Table 3, our method can achieve high accuracy on both the RAVDESS and EMO-DB datasets.

Table 3 Comparative analysis of the method proposed in this article with other methods

model	Accuracy (%)	
	RAVDESS	EMO-DB
Ref. [15]	75.28	92.71
Ref. [16]	90.60	93.30
Ref. [11]	92.91	89.15
Ref. [17]	75.28	92.71
Ref. [18]	87.32	89.30
This work	93.15	93.60

4 Conclusion

This paper proposes a speech emotion recognition model (TWM) based on improved MFCC features, fused multi-head attention mechanism, and improved Wasserstein Generative Adversarial Network (WGAN-GP). This article introduces the first and second derivatives of MFCC features to significantly improve the ability to capture dynamic information in speech signals, thereby enhancing the accuracy and robustness of emotion recognition. The TWM model based on TIM-NET network not only effectively alleviates the limitations of traditional models in processing long time series by combining multi-head attention mechanism and WGAN-GP, but also improves the model's generalization ability to diverse emotional data through data augmentation techniques. Future research can further explore the potential applications in other emotion recognition tasks and attempt to combine more advanced feature extraction techniques and model optimization strategies to achieve higher recognition accuracy and stronger model robustness.

References

[1] Jiao D N, He X, Xu J H, et al. Design of intelligent vehicle multimedia human-computer interaction system. IOP Conference Series: Materials Science and Engineering. Bristol: IOP Publishing, 2019, 563(5): 052029. DOI: 10.1088/1757-899X/563/5/052029.

[2] Tao J, Chen J, Li Y. Review on speech emotion recognition. Signal Processing, 2023, 39(4): 571-587. DOI: 10.16798/j.issn.1003-0530.2023.04.001.

[3] Liu Y R, Zhang X Y, Chen G J, et al. VMD improves the emotional speech feature extraction of GFCC. Journal of Xi'an University of Electronic Science and Technology, 2019, 46(5): 240-250.

[4] Kumaran U, Radha Rammohan S, Nagarajan S M, et al. Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep CRNN. International Journal of Speech Technology, 2021, 24: 303-314. DOI: 10.1007/s10772-020-09792-x.

[5] Cui L, Cui C L, Liu Z W, et al. Improved MFCC and parallel mixed model for speech emotion recognition. Computer Science, 2023, 50(6A): 220800211-7. DOI: 10.11896/jsjkx.220800211.

[6] Zhang L, Du J, Li J, et al. Speech emotion recognition method based on time-aware bidirectional multi-scale network. Proceedings of the 4th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA). Bristol: IOP Publishing, 2024, 2816(1): 012102. DOI: 10.1088/1742-6596/2816/1/012102.

[7] Fan Y H, Huang H M, Zhang H Y. Speech emotion recognition based on focal loss and ATCN-GRU. Computer Simulation, 2024, 41(2): 249-254+506.

[8] Zhang Y M, Qi H Y, Guo A. A student data generation model based on WGAN and multi-head attention mechanism. Journal of North Industrial University, 2024, 36(1): 76-83.

[9] Jiao Y M, Zhou C Z, Li W P. Speech emotion recognition research with multi-head attention fusion in VGGNet. Foreign Electronic Measurement Technology, 2022, 41(1): 63-69.

[10] Zhou J X, Jiao Y M, Wang Y B, et al. Speech emotion recognition research using attention and auxiliary classifier

in dilated residual network. Foreign Electronic Measurement Technology, 2023, 42(8): 19-25.
[11] Zhang K X, Liu Y X. Speech emotion recognition with multiple languages integration. Electronic Design Engineering, 2023, 31 (6): 25-29.
[12] Zhang X L. Improved MFCC features and MLA model for speech emotion recognition. Fujian Computer, 2024, 40(1): 52-56.
[13] Cui C L, Cui C. Lightweight speech emotion recognition aimed at data augmentation. Computer and Modernization, 2023, 2023(4): 83-89+100.
[14] Ye J, Wen X C, Wei Y, et al. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023, 1-5. DOI: 10.1109/ICASSP49357.2023.10096370.
[15] Zhao J F, Mao X, Chen L J. Speech emotion recognition using deep learning: DNN and 2D CNN-LSTM networks. Biomedical Signal Processing and Control, 2019, 47: 312-323. DOI: 10.1016/j.bspc.2018.08.035.
[16] Jahangir R, Teh Y W, Mujtaba G, et al. Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and feature fusion. Machine Vision and Applications, 2022, 33: article number 41. DOI: 10.1007/s00138-022-01294-x.
[17] García-Ordás M T, Alaiz-Moretón H, Benítez-Andrades J A, et al. Sentiment analysis in non-fixed length audios using a fully convolutional neural network. Biomedical Signal Processing and Control, 2021, 69: 102946. DOI: 10.1016/j.bspc.2021.102946.
[18] Zhou J, Liu J, Gan J, et al. A classroom speech emotion recognition method based on multi-scale temporal perception network. Computer Applications, 2024, 44(5): 1636-1643.