ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

Self-attentive Residual TCN Speech Emotion Recognition with Fused Acoustic Features

Bin Zhang, Yanping Zhu*

School of Microelectronics and Control Engineering, Changzhou University, Changzhou, China
**Corresponding author*

Abstract: *In order to improve the overall performance of the speech emotion recognition system, the problem of insufficient emotion information due to a single speech feature and the problem of recognition models not making full use of the emotion information contained in the features are addressed. In this paper, a self-attentive residual temporal convolution network (S-ResTCN) fusing Mel frequency cepstrum coefficients with rhythmic features is proposed. Firstly, the rhythmic features and mel frequency cepstral coefficients of speech were extracted on the EMO-DB and CASIA databases respectively, and their statistical functions were calculated to form 128-dimensional acoustic fusion features; then, the S-ResTCN network was designed and built, and the dependency modeling between the feature elements was completed by using the residual temporal convolution network, which made the network pay more attention to the parameters related to the emotional state in the features through the self-attentive mechanism, and generated the self-attentive mechanism feature matrix; finally, the softmax function was used for classification and recognition. The results showed that the S-ResTCN network improved the accuracy by 1.52%-14.12% over the existing network of the EMO-DB database and improved the accuracy by 1.27%-6.53% over the existing network of the CASIA database.*

Keywords: *speech emotion recognition, temporal convolution network, self-attention mechanism, mel-frequency cepstral coefficients, rhyme features*

1. Introduction

Speech emotion recognition (SER)[1]enables computers to predict the changing patterns of emotions carried in a speaker's speech by analyzing and processing the correlations between feature information in speech and emotions and is one of the important research directions in human-computer interaction. Acoustic features emotional information and highly accurate classification and recognition networks are important components of the SER task[2].

In recent years, researchers have proposed many speech features based on human voice characteristics, among which the mel-frequency cepstral coefficients (MFCC)[3], linear predictive coding (LPC)[4], and rhythmic features[5] are the most classic. The most classical methods are at the same time, researchers have proposed some more targeted features for the continuity and temporality of speech signals. For example, the literature[6] proposed non-linear geometric features and demonstrated that the proposed features could not only effectively characterize the emotional variability in speech signals, but also make up for the shortcomings of features in portraying emotional states. The literature[7] explored the relationship between the dimensional spatial model of emotion and speech features and extracted dimensional features corresponding to arousal and validity. Although there has been some research on features such as MFCC, LPC, and resonance peaks, they are still single-feature exploration, and there is a lack of research on fusing multiple types of acoustic features for emotion recognition.

Speech emotion recognition models mainly include the traditional gaussian mixture model (GMM)[8] and support vector machine (SVM)[9] models, as well as the latest TCN[10] and long short-term memory (LSTM)[11]. The literature[12] addresses the speaker discrepancy problem by forming spectral features into a channel feature input network, combining convolutional neural networks (CNN), bi-directional long short-term memory (BiLSTM), and attention mechanism to build the model, and proposing a method to assign channel weights using a deep residual shrinkage network. The literature [13] proposes a main and auxiliary network speech emotion recognition algorithm. Using BiLSTM as the main network and CNN-GAP as the auxiliary network, the extracted depth features are fused with features in a main and auxiliary network to solve the problem of unsatisfactory recognition results. In addition, inspired by the human attention mechanism, scholars have proposed various attention mechanisms for different problems in the

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

field of speech signal processing, such as the channel attention mechanism, the spatial attention mechanism, and the temporal attention mechanism. Although the aforementioned studies have achieved good results, they have not taken into account the dependence and variability of deep-seated intra-feature elements.

In summary, the current SER research faces two main problems: 1) it only considers a single type of speech feature and does not fully use the complementarity between acoustic features for the classification task; 2) it only models the acoustic features and emotional states of the speaker and fails to use the dependency relationships of the elements within the features to generate a depth feature map at a deeper level, and because the model uses the same weights for all the extracted depth features, it makes the classification recognition effect much less effective, i.e. it fails to adaptively reassign weights to the depth feature map.

To address the problem of single acoustic features, this paper proposes a 128-dimensional MFEZ feature set incorporating mel frequency cepstrum coefficient, fundamental frequency, energy, and zero crossing rate. To address the problem that the network fails to adaptively reassign weights to the depth features, this paper proposes a residual temporal convolutional network with a self-attention mechanism (S-ResTCN) by means of the residual temporal convolutional network (ResTCN) that can be used to model the dependency analysis of the elements within the features, which in turn enables the model to take into account the correlation between the elements within the features when generating the depth feature map. The self-attention mechanism enables the calculation of the sentiment information stored in each feature in the depth feature map and the reallocation of the weight of each feature according to the proportion of sentiment information, so as to maximise the use of sentiment information in the features and improve the recognition rate of sentiment classification.

2. Self-attentive residual time convolutional networks

The S-ResTCN network consists of two parts, the ResTCN network, and the self-attentive mechanism. Among them, the ResTCN network part will complete the modeling of the correlation between MFEZ features and sentiment, while the self-attentive mechanism layer will redistribute weights according to the proportion of sentiment information carried by the features. As shown in Figure 1, the ResTCN network consists of a TCN and a residual connection, where the TCN network learns the sentiment information by processing the feature sequences in parallel, and then uses the residual connection to make the model generate a stable gradient optimization path during the training process; the Query and Key in the self-attentive mechanism are multiplied to obtain the sentiment information weight feature, and this is applied to the Value so as to increase the local The self-attentive weight feature is generated by multiplying Query and Key in the self-attentive mechanism. Finally, the softmax function is used for sentiment classification.

2.1. ResTCN

In traditional speech emotion recognition research, scholars tend to model only speaker features and emotion states to analyze correlations, while neglecting the analysis of correlations between elements within features, so this paper proposes a ResTCN network. ResTCN network consists of two parts, TCN, and residual connectivity, the TCN layer uses an inflated convolutional layer to process feature sequences in parallel. The expansion factor is set to model the intrinsic dependencies between elements at different positions in the feature sequence, and the residual connectivity will enable the model to produce more stable gradient optimization paths during training.

Figure 2 shows the principle of the i layer of the ResTCN network. Firstly, the input features x_i will pass through the Batch Normalization, ELU activation function, and Dropout layer (dropout=0.5) after passing through the dilated convolutional layer; then the output of the Dropout layer will be fused with x_i to obtain x_{i+1} and fed into the $i+1$ layer ResTCN; finally, the output of the last ResTCN layer will be fed directly to the Self-Attention Mechanism module.

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

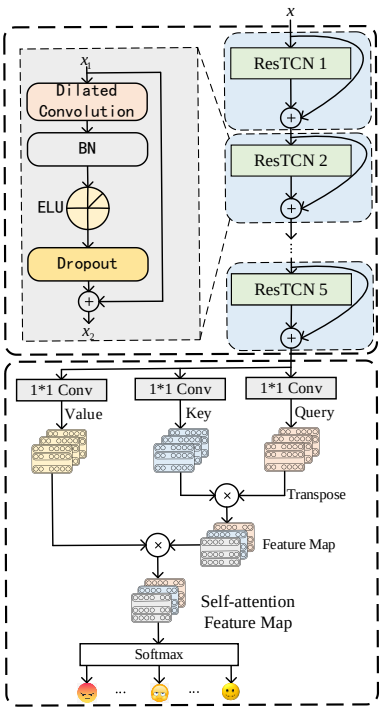


Figure 1: Block diagram of S-ResTCN-based speech emotion recognition network

The ELU activation function accelerates the learning of the mean towards zero by reducing the effect of the bias offset so that the normal gradient is closer to the unit's natural gradient. The ELU expression is shown in equation (1).

$$\mathrm{ELU}(x) = \begin{cases} x, x > 0 \\ \alpha\left(e^{x} - 1\right), x \le 0 \end{cases} \tag{1}$$

α is an adjustable parameter that controls the degree of saturation of the ELU in the negative part. When the input is greater than 0, the output can effectively alleviate the gradient disappearance; when the input is less than 0, the output means is as close to 0 as possible to improve the convergence speed.

The principle formula for residual connectivity in ResTCN networks is

$$\boldsymbol{x}_{i+1} = h\left(\boldsymbol{x}_i\right) + \Gamma\left(\boldsymbol{x}_i, w_{(i,d)}\right) \tag{2}$$

$i \in \{1,2,3,4,5\}$, $h(x_i)$ is the direct mapping part of x_i; $\Gamma\left(x_i, w_{(i,d)}\right)$ is the residual part, $w_{(i,d)}$ denotes the i layer convolution operation, and $d \in \{2^0, 2^1, 2^2, 2^3, 2^4\}$ is the expansion rate. In the training process, the convolution kernel of the inflated convolution layer is 2, and the number of filters is $f = \{25,26,27,28\}$ respectively.

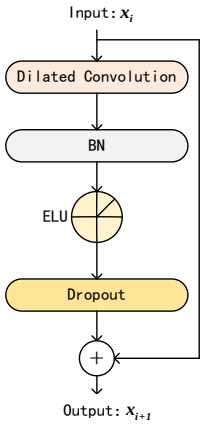


Figure 2: Layer I ResTCN network structure

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

2.2. Self-attention mechanism

Different types of feature parameters have different roles in sentiment recognition, but during model training, the network assigns the same weights to the feature maps, which will lead to the under-utilization of sentiment information. In this paper, we combine the ResTCN model with the self-attention mechanism to better utilize the information significantly related to emotion in the acoustic signal by weighting the output feature maps of the inflated convolutional layer in ResTCN with the emotion information.

The core module of the self-attention mechanism, which can better focus on the dependencies between the input features, is the scaled point multiplied attention mechanism. The scaled dot product attention mechanism non-linearly maps the elements in the input feature sequence to generate three different representations, Query, Key, and Value, where Query and Key denote the vectors for calculating the attention weights and Value denotes the input feature vector. Where, q_t, k_t, v_t are calculated as shown in equation (3).

$$\boldsymbol{q}_t = \boldsymbol{w}_q^T \boldsymbol{x}_t ; \boldsymbol{v}_t = \boldsymbol{w}_v^T \boldsymbol{x}_t ; \boldsymbol{k}_t = \boldsymbol{w}_k^T \boldsymbol{x}_t \tag{3}$$

w_q^T, w_v^T, w_k^T are the hyperparameters of the network, which are obtained by training and learning. In this paper, q_t , k_t, v_t of all elements of the input, features are represented by matrices as Q , K, and V. The dot product attention mechanism is calculated as shown in equation (4).

$$\boldsymbol{Z} = Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{4}$$

In the network, the features of speech are input as a matrix, and self-attention calculates the correlation coefficient between the current speech feature and other speech features by calculating the feature weights[13]. The process can be summarised in three stages, with the specific operations of each stage shown in Figure 3.

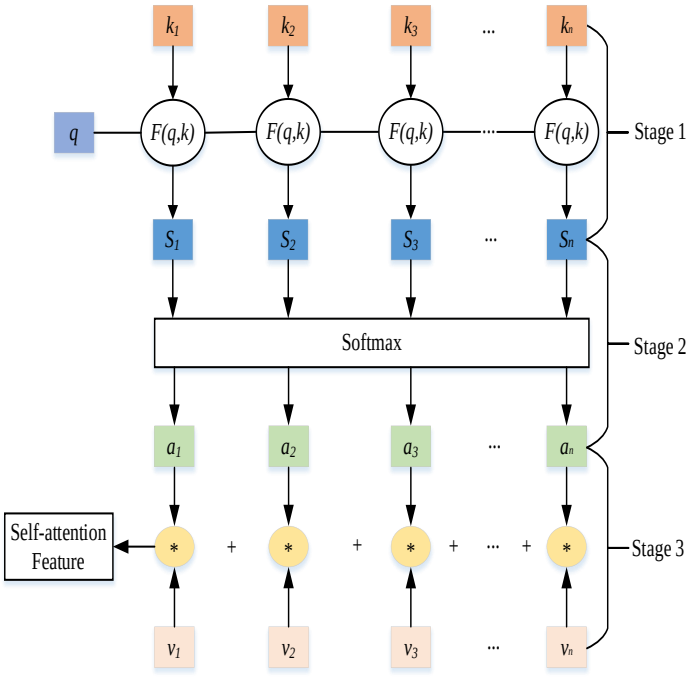


Figure 3: The calculation process of Self-attention

At stage 1, the dot product operation between q_i of group i and each group k is performed to obtain the similarity score matrix s_i. At stage 2, s_i is normalized by the softmax function to ensure that the weight parameters of s_i sum to 1. At stage 3, the resulting matrix of weight coefficients $\{a_1, a_2, \cdots, a_n\}$ is applied to v , which is weighted and summed to obtain the final self-attentive feature matrix.

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

3. Fusion of rhythmic features and MFCC features

3.1. Rhythmic features

Rhyme features reflect changes in the intensity and intonation of speech emotion signals. The rhyme features extracted in this paper are shown below.

1) Zero crossing rate: The number of times a waveform crosses the zero level in a frame of speech is known as the zero crossing rate, which is defined as shown in equation (5).

$$\boldsymbol{Z}_n = \frac{1}{2}\sum_{m=0}^{N-1}\left|\mathrm{sgn}\left[\boldsymbol{x}_n(m)\right]-\mathrm{sgn}[\boldsymbol{x}_n(m-1)]\right| \tag{5}$$

2) Energy: Let the short-time energy of the n frame of the speech signal be expressed as

$$\boldsymbol{E}_n = \sum_{m=0}^{N-1}\boldsymbol{x}_n^2(m) \tag{6}$$

3) Fundamental frequency: The frequency at which the vocal folds vibrate when a turbid tone is pronounced. When a person pronounces a sound, the vocal tract is strongly excited, and this is reflected in a dramatic increase in the amplitude of the speech waveform. The inverse of the length of time between the closure of two adjacent vocal folds is the fundamental frequency at that point.

3.2. MFCC features

Mel frequency is a speech characteristic parameter constructed from the auditory properties of the human ear. Since the sound height heard by the human ear does not correspond linearly to frequency, but more closely to a logarithmic relationship, the mayer frequency scale accurately corresponds to the auditory characteristics of the human ear, and its relationship with frequency can be expressed as

$$F_{mel} = 2595\lg(1+ f_{hz}/700) \tag{7}$$

The steps to extract the MFCC are as follows: 1) first make the signal pass through a high-pass filter for pre-emphasis; 2) perform frame-splitting and windowing; 3) perform a fast fourier transform to obtain the spectrum of each frame; 4) pass the power spectrum through a set of meier-scale triangular filter banks with a filter order of 24 and then log the result; 5) finally, after a discrete cosine transformer, the MFCC coefficients can be obtained.

3.3. MFEZ feature set

In this paper, the over-zero rate, energy, fundamental frequency, and the corresponding first-order difference coefficients of the rhythmic features are extracted, and then their statistical functions are calculated and fused with the MFCC statistical functions to form the 128-dimensional MFEZ features, which are represented as shown in Equation (8).

$$F_u = \left\{Z, \Delta Z, E, \Delta E, F, \Delta F, M_1, M_2, \cdots, M_k\right\} \tag{8}$$

where: Z and △Z denote the vector of statistical functions for the low-level descriptors of the over-zero rate and the first-order difference coefficients, respectively; E and △E denote the vector of statistical functions for the low-level descriptors of the energy and the first-order difference coefficients, respectively; F and △F denote the vector of statistical functions for the low-level descriptors of the fundamental frequency and the first-order difference coefficients, respectively, where the vector equation for Z is shown in equation (10), and the vector equations for Eand F are simply substituted for the variable Z in equation (9).

$$Z = \{max(z), min(z), ran(z), maxpos(z), minpos(z), \bar{z}, lin(z), std(z)\} \tag{9}$$

The variables in Eq. are, in order, the maximum value, minimum value, range value, absolute range of the maximum value, absolute range of the minimum value, mean value, slope, and standard deviation value of Z. In addition, m in equation (8) denotes the vector of statistical functions constituting the k order MFCC, as shown in equation (10).

$$M_k = \{max(M_k), min(M_k), ran(M_k), maxpos(M_k), minpos(M_k), \overline{M_k}, lin(M_k), std(M_k)\} \tag{10}$$

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

The variables in the equation are, in order, the maximum value, minimum value, range value, absolute range of the maximum value, absolute range of the minimum value, mean value, slope, and standard deviation value of the k order MFCC. In this paper, k taking a value of 10, the final 128-dimensional fused acoustic features that make up the MFEZ feature set are shown in Table 1 with the following specific information.

Table 1: MFEZ feature set details

Features	Description of features	Dimensionality
E & ΔE	Energy and first-order difference coefficients	16
Z & ΔZ	Trans-zero rates and first-order difference coefficients	16
F & ΔF	Fundamental frequency and first-order differential coefficients	16
M_k	MFCC and kth order difference coefficient	80

4. Experimental design and analysis of results

4.1. Speech emotion database

In this paper, the EMO-DB databases and the CASIA databases are used, and the specific information of the databases is shown in Table 2.

Table 2: Database details

Databases	Language	Emotional category	Emotional state	Number
EMO-DB	German	7	Anger	71
			Sadness	69
			Happy	81
			Scared	46
			Neutral	127
			Disgusted	79
			Bored	62
CASIA	Chinese	6	Anger	200
			Surprise	200
			Fear	200
			Happy	200
			Jealous	200
			Sadness	200

In Table 2, the EMO-DB was recorded at the Technical University of Berlin, with ten subjects simulating German speech for each of the seven emotions, emotions including anger, sad, happy, scared, neutral, disgusted, and bored, with a total of 535 speech data. CASIA was designed and recorded at the Institute of Automation, Chinese Academy of Sciences, by four subjects in a noiseless environment, with a male-to-female ratio of 1:1, with each corpus data expressing six emotions: anger, surprise, fear, happiness, jealousy, and sadness, with a total of 1200 speech data.

4.2. Experimental environment settings and evaluation indicators

Software environment: Windows 10 as the operating system, tensorflow 2.3.0 as the deep learning framework, python 3.6 as the programming language environment, and NVIDIA GTX 1080Ti as the GPU.

Model training method and parameter settings: S-ResTCN sentiment recognition network randomly divided the dataset, the training set, validation set, and test set was divided in the ratio of 6:2:2. Adam was chosen as the optimizer, the learning rate was set to 0.001, the decay index of the first estimation was set to 0.5, and the batch size was set to 128.

To validate the effectiveness of the S-ResTCN network, Accuracy, Area Under The Curve(AUC), and Confusion Matrix were used as evaluation metrics.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

$$\mathrm{AUC} = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(1+\mathrm{M})}{2}}{M \times N} \tag{12}$$

Eq. (11) represents the proportion of samples correctly classified in the total sample, where TP is the true case, TN is the true negative case, FP is the false positive case and FN is the false negative case. Eq. (12) represents the probability that any positive sample is greater than a negative sample, where M is the number of positive samples, N is the number of negative samples, and $\mathrm{rank_i}$ represents the number of the i sample.

4.3. Experimental results and analysis

4.3.1. S-ResTCN ablation experiment

To demonstrate the role of S-ResTCN in sentiment recognition and to explore the effectiveness of ResTCN and Self-attention in sentiment recognition, this subsection compares the sentiment classification recognition rates of TCN, ResTCN, and S-ResTCN using 128-dimensional MFEZ features from the EMO-DB and CASIA databases as input. The results of the ablation experiments are shown in Table 3.

Table 3: Ablation experiments

Database	Model	Accuracy	AUC
EMO-DB	TCN	82.09%	82.63%
	ResTCN	89.64%	89.78%
	S-ResTCN	96.62%	97.13%
CASIA	TCN	80.36%	80.38%
	ResTCN	87.17%	87.61%
	S-ResTCN	92.35%	93.45%

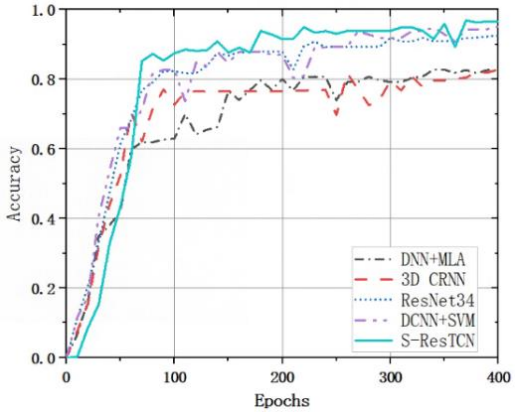
Comparing Table 3, it can be found that the ResTCN network improved the sentiment recognition accuracy by 7.55% and 6.81% on EMO-DB and CASIA, respectively, compared to the TCN network, proving that the residual connection is beneficial to improve the sentiment recognition accuracy. Comparing the S-ResTCN with the TCN and ResTCN networks respectively, the accuracy improvement was 14.53% and 6.98% on the EMO-DB database and 11.99% and 5.18% on the CASIA database. In addition, the S-ResTCN also achieved a certain degree of improvement over the TCN and ResTCN networks in terms of the AUC metric, indicating that the self-attentive mechanism can effectively re-weight the feature maps output by the ResTCN, allowing the sentiment information in the MFEZ features to be more fully explored.

4.3.2. S-ResTCN versus other networks

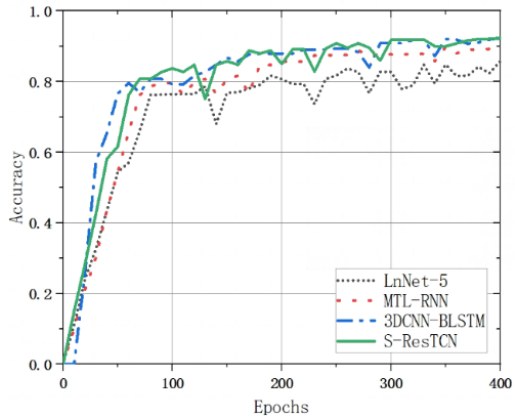
In order to further verify the effectiveness of the proposed network in sentiment recognition, the MEEZ features of the EMO-DB and CASIA databases were extracted as input, and the S-ResTCN network was compared with the existing networks of the EMO-DB and CASIA databases respectively. Among them, the existing networks compared in the EMO-DB database are DNN+MLA, 3D CRNN, ResNet34, and DCNN+SVM; the existing networks compared in the CASIA database are LnNet-5, MTL-RNN, and 3DCNN-BiLSTM, and the comparison results are shown in Figure 4.

From Fig. 4(a), it can be seen that S-ResTCN has the worst accuracy and poor model stability in comparison with the existing models of the EMO-DB database, with the best performance stabilized at around 82.50% for DNN+MLA; the best performance of 3D CRNN network is around 82.82%; the best performance of ResNet34 and DCNN+SVM networks are stabilized at 92.41% and 95.10%; the accuracy of S-ResTCN network was around 96.62%, possessing a higher accuracy rate compared to other classification networks. As shown in Fig. 4(b), the accuracy of S-ResTCN was the worst in comparison with the existing models in the CASIA database, LnNet-5, which finally stabilized at around 85.82%; the accuracy of MTL-RNN and 3DCNN-BiLSTM networks were 90.91% and 91.08% respectively; the accuracy of S-ResTCN network was 92.35%, comparing with the S-ResTCN network has better recognition results than the current existing networks.

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306



(a) Comparison of different networks of the EMO-DB database with the S-ResTCN network



(b) Comparison of different networks of the CASIA database with the S-ResTCN network

Figure 4: Comparison of the S-ResTCN network with existing networks of different databases

The experimental results of the proposed method under the EMO-DB and CASIA databases are compared with the results of the existing literature as shown in Table 4. By comparing Table 4, it can be found that S-ResTCN outperforms the comparative literature in classification recognition on the EMO-DB and CASIA databases. In the EMO-DB database comparison analysis, S-ResTCN improved the accuracy by 14.12%, 13.8%, 4.21%, and 1.52%, and the AUC by 14.17%, 14.02%, 4.26%, and 1.20% over DNN+MLA, 3D CRNN, ResNet34, and DCNN+SVM, respectively. In the CASIA database comparison analysis, S-ResTCN improved the accuracy over LeNet, MTL-RNN, and 3DCNN-BLSTM by 6.53%, 1.44%, and 1.27%, and the AUC by 6.61%, 1.9% and 1.6%, respectively.

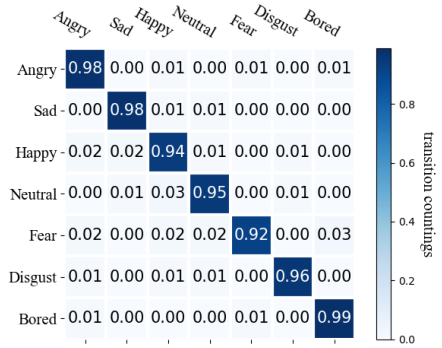
Table 4: Comparison of S-ResTCN with other networks

Database	Model	Accuracy	AUC
EMO-DB	DNN+MLA	82.50%	82.96%
	3D CRNN	82.82%	83.11%
	ResNet34	92.41%	92.87%
	DCNN+SVM	95.10%	95.93%
	S-ResTCN(our)	**96.62%**	**97.13%**
CASIA	LeNet-5	85.82%	86.84%
	MTL-RNN	90.91%	91.55%
	3DCNN-BLSTM	91.08%	91.85%
	S-ResTCN(our)	**92.35%**	**93.45%**

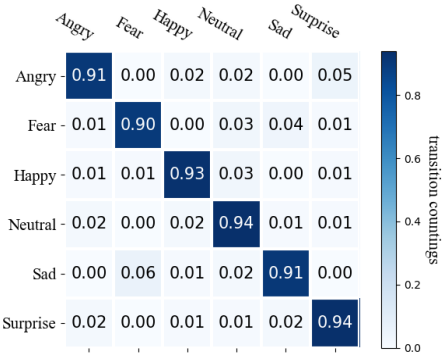
Since the classification accuracy score refers to the percentage of all correct classifications but does not reveal the potential distribution of response values, so to verify the reliability of the classification accuracy, the confusion matrix is chosen to calculate the classification accuracy under different sentiment states in this paper, and the results are shown in Figure 5. By comparing Fig. 5(a), it can be found that the accuracy of happy and scared in S-ResTCN under the EMO-DB database is slightly lower, reaching 94% and 92%, while the accuracy of other emotions is relatively stable, between 96% and 99% range. By comparing Figure 5(b), it can be concluded that in the CASIA database, the accuracy rates of the emotional states of fear and sadness are lower, and the accuracy rates of the neutral and surprised states

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

are the highest, reaching 94%. It further proves that the S-ResTCN network proposed in this paper can better process the emotion information in speech features and improve speech emotion recognition accuracy.



(a) Confusion matrix under EMO-DB database



(b) Confusion matrix under CASIA database

Figure 5: S-ResTCN network obfuscation matrix

5. Conclusion

In this paper, the 128-dimensional MFEZ fusion acoustic feature set is proposed by combining the complementary advantages of rhyme features in terms of over-zero rate, energy, fundamental frequency, and MFCC features in speech emotion recognition; to improve the speech emotion recognition accuracy, an S-ResTCN network is designed and built. The model introduces the correlation between feature elements and adaptively reassigns weights to the depth feature map based on the differences in the emotional information carried by different types of features, so that the generated depth feature map incorporates both the internal dependency relationship between feature elements and the differences in the emotion contained in the features, thus maximizing the analysis and modeling of the emotional elements in acoustic features.

Acknowledgement

Fund project: 1) Changzhou Key Research and Development Program (No. CJ20210123); 2) Jiangsu Postgraduate Research Innovation Project (No. KYCX22_3053).

References

[1] Khalil R A, Jones E, Babar M I, et al. Speech emotion recognition using deep learning techniques: A review [J]. IEEE Access, 2019, 7: 117327-117345.
[2] Milton A, Roy S S, Selvi S T. SVM scheme for speech emotion recognition using MFCC feature[J]. International Journal of Computer Applications, 2013, 69(9):183-194.
[3] Sun L, Zou B, Fu S, et al. Speech emotion recognition based on DNN-decision tree SVM model[J]. Speech Communication, 2019, 115: 29-37.
[4] Wang W, Watters P A, Cao X, et al. Significance of phonological features in speech emotion

ISSN 2616-5775 Vol. 6, Issue 3: 42-51, DOI: 10.25236/AJCIS.2023.060306

recognition[J]. International Journal of Speech Technology, 2020, 23: 633-642.
[5] Han D, Kong Y, Han J, et al. A survey of music emotion recognition [J]. Frontiers of Computer Science, 2022, 16(6): 166335.
[6] Yi Y, Tian Y, He C, et al. DBT: multimodal emotion recognition based on dual-branch transformer[J]. The Journal of Supercomputing, 2022: 1-23.
[7] SHARMA P, ABROL V, DILEEP A, et al. Class specific GMM based sparse feature for speech units classification[C]. European Signal Processing Conference. SHARMAP, 2017.
[8] SCHULLER B, RIGOLL G, LANG M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture[C]. IEEE international conference on acoustics, speech, and signal processing. SCHULLERB, 2004.
[9] Ye Y, Chen J. Multi-modal Speech Emotion Recognition Based on TCN and Attention[C]. Proceedings of the 11th International Conference on Computer Engineering and Networks. Springer Singapore, 2022.
[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[11] Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical signal processing and control, 2019, 47: 312-323.
[12] Dangol R, Alsadoon A, Prasad P W C, et al. Speech emotion recognition Using Convolutional neural network and long-short Term Memory[J]. Multimedia Tools and Applications, 2020, 79: 32917-32934.
[13] Desheng H, Xueying Zhang, et al. Speech emotion recognition based on primary and secondary network feature fusion[J]. Journal of Taiyuan University of Technology, 2021, 52(05):769-774.