Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 10, No. 1, 2023

Speech emotion recognition based on dynamic convolutional neural network

Ziyao Lin[1,2,*], Zhangfang Hu[1], Kuilin Zhu[1,2]

[1] Key Laboratory of Optoelectronic Information Sensing and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] School of Advanced Manufacturing Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[*] **Corresponding author**: Ziyao Lin

Abstract: In speech emotion recognition, the use of deep learning algorithms that extract and classify features of audio emotion samples usually requires the use of a large amount of resources, which makes the system more complex. This paper proposes a speech emotion recognition system based on dynamic convolutional neural network combined with bi-directional long and short-term memory network. On the one hand, the dynamic convolutional kernel allows the neural network to extract global dynamic emotion information, which can improve the performance while ensuring the computational power of the model, and on the other hand, the bi-directional long and short-term memory network enables the model to classify the emotion features more effectively with the temporal information. In this paper, we use CISIA Chinese speech emotion dataset, EMO-DB German emotion corpus and IEMOCAP English corpus to conduct experiments, and the average emotion recognition accuracy of the experimental results are 59.08%, 89.29% and 71.25%, which are 1.17%, 1.36% and 2.97% higher than the accuracy of speech emotion recognition systems using mainstream models, respectively. The effectiveness of the method in this paper is proved.

Keywords: Speech emotion recognition; Dynamic convolutional kernel; Neural network.

1. Introduction

Speech is a complex signal that contains a lot of information such as speaker semantics and emotion and uses language as an information carrier [1]. Speech emotion recognition technology is to extract the features in the speech signal that can characterize the speaker's emotional state, and find out the mapping relationship between these features and human emotion through machine learning and other methods [2]. The ultimate goal is to enable machines to recognize the emotional state of a speaker and achieve the goal of intelligent and harmonious human-computer interaction. Speech emotion recognition has been around for decades, and with the development of deep learning in recent years, there are more new developments in speech emotion recognition technology, which has great promise for applications in human-computer interaction, in-car navigation, in the field of teaching aids, in medical therapy, in robotics, and even in the field of video games [3]-[6]. Therefore, the research of speech emotion recognition technology is very meaningful and has high application value.

Speech is a continuous signal of varying length that carries the message and emotion expressed by the speaker. Emotions consist of several different signals such as happy, angry, sad, calm, bored, disgusted and fearful [7]. The emotional features in speech signals can be extracted by the classification model. Speech emotional features are divided into three main categories: rhythmic features, phonological-related features, and spectral features [8]-[9].

Rhythmic features are those features that can be perceived by humans, such as intonation and rhythm, which are the most significant features for expressing emotional content in speech emotion recognition [10]-[14]. Sound quality features are used to measure whether the speech is pure, clear, easily recognizable, etc. Spectral features are a feature of the relationship between vocal tract shape variation and speaker vocalization [15].

The most popular algorithms for traditional speech emotion recognition systems are Hidden Markov Models (HMM)[16], Gaussian Mixture Models (GMM)[17], Support Vector Machines (SVM)[18], and Artificial Neural Networks (ANN)[19]. There are also algorithmic approaches based on decision trees (DT) [20], k-nearest neighbors (KNN) [21], etc. Among them, HMM is suitable for the recognition of time-series sequences, but is influenced by the phonemic information; GMM has a better ability to fit the data, but is highly dependent on the training data; SVM is suitable for small sample training sets, but is deficient in multi-class classification problems; ANN has the advantage that it approximates complex nonlinear relationships, but it is easy to fall into local minimal features and the algorithm convergence speed is slow.

In recent years, deep learning algorithms have outperformed traditional machine learning algorithms, so the focus of research has shifted to them, and the trend in current research on speech emotion recognition is the same. The most widely used deep learning algorithms in the field of speech emotion recognition are convolutional neural networks (CNN) and recurrent neural networks (RNN).

Convolutional neural networks are a special type of neural network designed to process data with a lattice-like topology, such as images as well as speech features in two-dimensional space [22]. By applying several correlation filters, convolutional neural networks can effectively capture temporal and spatial dependencies from one input source. Reducing the input to one form without losing features reduces the computational complexity and improves the success of the algorithm [23]. However, traditional modern high-performance convolutional neural networks often require a large amount of computational resources to perform a large number of convolutional kernel operations, and the nature of traditional convolution causes a large amount of

redundant computations inherent to convolutional neural networks.

RNN is able to process the whole time series information, but its memory is most profound for some last input signals, while the strength of signals earlier will get lower and lower, and finally can only play some auxiliary role, so the recurrent neural network only has short term memory, but by using long-short term memory structure, RNN can access long term memory. Long-short-term memory network (LSTM) is a class of gated recurrent neural networks, which is the most effective model for practical application to solve the long-term dependence problem of RNN, and is also widely used in speech emotion recognition [24]. However, the long short-term memory network can only extract the previous information in a one-way manner, and cannot fully consider the rich emotional information contained in complex human language in its context, ignoring the influence of the later text on the previous information.

To address the above problems, this paper proposes a speech emotion recognition model based on DyCNN(dynamic convolutional neural network)[25]-[28] and Bi-LSTM. Firstly, the performance of the algorithm is improved by dynamic convolutional method, which reduces the redundant computation of the network, improves the flexibility of static convolutional kernel for different emotion information extraction, extracts the global emotion information, and combines the attention mechanism to assign different attention weight values to different feature regions in speech, which will better extract the features of the prominent part of emotion role in a sentence of speech. Combined with Bi-LSTM, it solves the problem of long-term dependence of traditional RNN and the problem of insufficient extraction of contextual information by LSTM, and makes more efficient use of temporal information. The simulation results show that the model proposed in this paper can effectively improve the recognition rate of speech emotion recognition system.

2. DyCNN model combining Bi-LSTM

2.1. Model Framework

In this paper, we use a combination of a LSTM and a DyCNN as the network architecture, which is shown in Fig 1. The model first preprocesses the initial speech signal to obtain the log-mel spectrogram, then inputs the global dynamic spectrogram features into the DyCNN to extract the global dynamic spectrogram features, then inputs the Bi-LSTM to extract the temporal sentiment information combined with the context, and finally uses the Softmax layer for sentiment classification. Each module is described as follows.

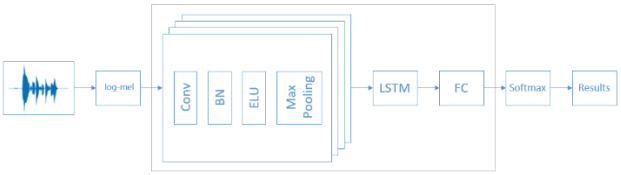

Fig.1 The network architecture

2.2. Feature extraction

Log-mel spectrogram is a method that can reflect the information of mood change from the speech signal, so it is used as the input of the network. In this paper, the original speech signal is firstly obtained, pre-emphasized, framed, windowed, and short-time Fourier transformed, and then the obtained acoustic spectrogram is input to the mel filter set to obtain the mel spectrogram, and then the log-mel spectrogram is obtained by taking the logarithm of it, and finally the spectrogram features are input to the subsequent network.

2.3. Dynamic Convolutional Neural Network

The static convolutional kernel used in traditional convolutional neural networks shares its parameters for different input samples, while for speech emotion recognition, it is obvious that different speakers and different contents are more advantageous to use dynamic convolution than static convolution. Therefore, in this paper, we use a dynamic convolutional neural network capable of adaptively changing attention according to the input to build a speech emotion recognition system.

This paper uses dynamic convolution to improve the convolutional part of the baseline network architecture. Instead of using a single convolutional kernel at each layer, dynamic convolution dynamically aggregates multiple parallel convolutional kernels based on their attention mechanisms, which are input-dependent. This set of convolutional kernels is weighted and summed with a matrix of preceding attention weight parameters to obtain dynamic convolutional kernels that can adaptively change their attention according to the input, and convolve with the input spectral map to obtain more emotionally informative features. A dynamic convolutional layer is shown in Figure 2 below.

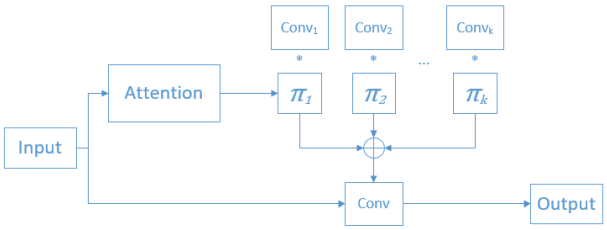

Fig.2 A dynamic convolutional layer

The attention mechanism uses an average pooling layer and two fully connected layers with low and efficient model complexity, and Softmax is used to restrict the attention weights πk to between 0 and 1, enabling the model to learn features in a deep level. For the feature map xi generated in the convolution process, it is first operated several times to generate K attention weight parameters πk that sum to 1, and then the K convolution kernel parameters are linearly summed, so that the convolution kernel obtained during inference is changed with the transformation of the input.

The dynamic convolution kernel model is calculated as follows.

$$y = \sigma\left(\sum_{k=1}^{K} \pi_k(x)\widetilde{W}_k \cdot x + \sum_{k=1}^{K} \pi_k(x)\tilde{b}_k\right) \quad (1)$$

where y is the output, x is the input, σ is the ELU activation function, πk is the attention weight, $\widetilde{W}_k$ is the weight matrix, and $\tilde{b}_k$ is the bias term. The formula shows that x performs two operations, one for finding the parameters of attention to generate the dynamic convolution kernel and one for being convolved.

2.4. Bidirectional long- and short-term memory network

The LSTM model is based on the RNN with the addition of input gates, forgetting gates, unit states, and output gates.

During network training, information can be added or removed through the gate structure, and different neural networks can decide which relevant information to remember or forget through the gate structure on the unit state. At moment t, each gate state update can be expressed as follow.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1} \cdot x_t] + b_i) \quad (3)$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * tanh(W_C \cdot [h_{t-1} \cdot x_t] + b_C) \quad (5)$$

$$h_t = o_t * tanh(C_t) \quad (6)$$

where ft is the forgetting gate, it is the input gate, ot is the output gate, Ct is the cell storage cell, W* is the weight matrix, xt and ht are the input vector and the hidden layer vector at the t-th time step, respectively, the input vector, b* is the bias term, and σ is the activation function.

In this paper, we use Bi-LSTM to replace the LSTM part of the baseline network architecture. bi-LSTM is a combination of forward LSTM and backward LSTM. Like LSTM, Bi-LSTM is often used to model contextual information in natural language processing tasks. However, it is impossible to encode back-to-front information when modeling speech emotion signals using LSTM, while Bi-LSTM can solve this problem well. The fig 3 shows the structure of Bi-LSTM expanded along time.

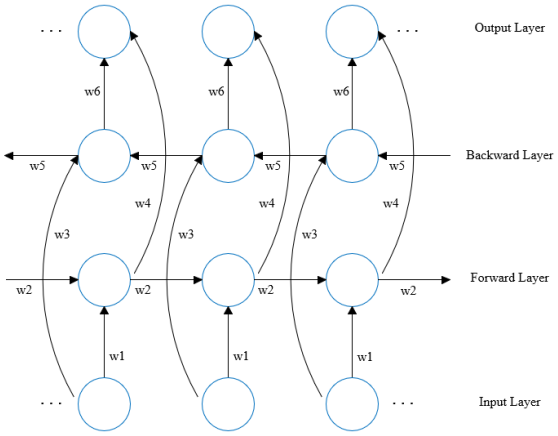

Fig.3 The structure of Bi-LSTM expanded along time

As shown in the figure, the Bi-LSTM consists of two one-way LSTMs, the forward LSTM and the backward LSTM, the former is responsible for calculating the forward context information and the latter is responsible for calculating the backward context information, thus, the Bi-LSTM allows the model to obtain the complete context information and thus the recognition accuracy is improved.

3. Experiments

3.1. Speech emotion dataset

Speech emotion recognition is to identify the emotional state of a speaker when communicating by voice. Natural speech in daily life has rich and variable emotions and the actual environment is extremely complex, so it is extremely complex and difficult to capture natural speech in real life to create a speech emotion dataset, which needs to meet four conditions such as real, continuous, interactive, and diverse and exclude external interference as much as possible. Therefore, speech emotion recognition research generally uses a corpus of speech emotion datasets recorded in artificially provided quiet environments such as recording studios.

Speech emotion corpus is the basis for speech emotion recognition, and a large, diverse, high-quality corpus is essential for the performance of speech emotion recognition systems. Among them, discrete emotion datasets have relatively single emotion state, easy to recognize several simple emotion speech signals, the following is an introduction to several common speech emotion corpora used for speech emotion classification and recognition of public discrete emotion.

1) The EmoDB corpus is a discrete German emotion corpus derived from the Berlin laboratory. The corpus was recorded by 5 men and 5 women in the form of performances with 535 sentences containing 7 types of emotions: happy, angry, sad, calm, bored, disgusted and fearful.

2) The CASIA corpus is a discrete Chinese emotion corpus recorded by the Institute of Automation, Chinese Academy of Sciences. The corpus is recorded by 2 male and 2 female 4 professional pronouncers in the form of performance, containing 6 emotion types: angry, afraid, happy, neutral, sad and surprised, with 9600 different pronunciations, containing 300 identical texts and 100 different texts.

3) The IEMOCAP corpus was collected and recorded by the Sail Lab at the University of Southern California and is one of the largest dimensional speech emotion databases for speech emotion recognition. The corpus consists of approximately 12 hours of spoken dialogue in the form of conversational performances by two of the 10 speakers in five separate stages. The corpus was annotated by at least two annotators with labels such as anger, happiness, sadness, and neutrality, as well as three emotion dimensions: arousal dimension, valence dimension, and dominance dimension.

Experimental setup and evaluation index

In order to verify the effectiveness of the model proposed in this paper on speech emotion recognition system, the above three corpora are chosen to experiment on the model, and the weighted average accuracy (WA) is used as the experimental evaluation index, and compared with the existing mainstream models with the following equation, where n denotes the number of correctly identified test samples and N denotes the total number of test samples.

$$WA = \frac{\mathrm{n}}{\mathrm{N}} \quad (7)$$

The experimental environment was conducted under Windows with AMD Ryzen 9 4900H processor, based on Pytorch framework, and trained on NVIDIA GeForce RTX 2060 graphics card. In the experiments, the speech signal is sampled and then pre-emphasized by a first-order FIR digital high-pass filter with a pre-emphasis factor of 0.97, where the frame length is taken as 30ms, the frame shift is taken as 15ms, and the frame splitting is done using a Hamming window with the following equation, where $S_w(n)$ is the windowed speech signal, $S(n)$ is the input speech signal, and $w(n)$ is the window function.

$$S_w(n) = S(n)w(n) \quad (8)$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)), & 0 \le n \le N-1 \\ 0, & others \end{cases} \quad (9)$$

Then MFCC feature parameter extraction is performed, whose data dimension is 39 dimensions, and the extracted speech spectrogram is normalized, and the result is used as the input data of the model.

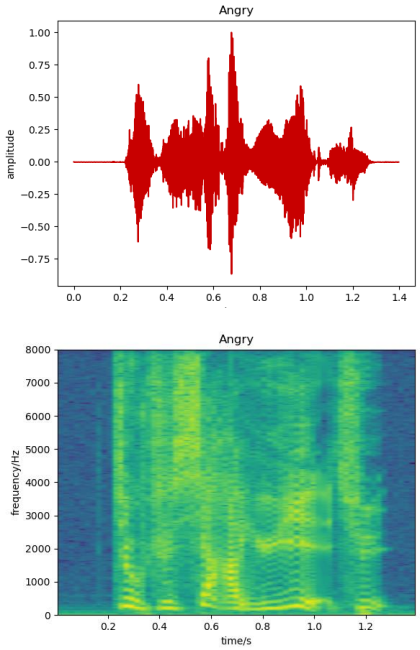

Fig.4 Original speech signal and speech spectrogram of the anger sample

In this paper, the TensorFlow toolkit is used to build the network model and complete the implementation of the training algorithm, and the model parameters are optimized by the RMSProp algorithm, the initial learning rate is set to 0.01, the cross entropy is used as the loss function, the training batch is 200, and the number of iterations is 1000.

3.2. Results and Analysis

In this paper, a baseline network is set up in the experiment to verify the effectiveness of the proposed algorithm, and the baseline network in this paper is a network model with CNN+LSTM serial setup, and the detailed data comparison is carried out in Table 1 below.

Table 1. Comparison with baseline network

Database	Model	WA
CASIA	CNN+LSTM	57.91%
CASIA	DYCNN+Bi-LSTM	59.08%
EmoDB	CNN+LSTM	87.93%
EmoDB	DYCNN+Bi-LSTM	89.29%
IEMOCAP	CNN+LSTM	68.28%
IEMOCAP	DYCNN+Bi-LSTM	71.25%

As can be seen from Table 1, the WA of the proposed network model in this paper is 59.08%, 89.29% and 71.25% on the three datasets, respectively, which is a certain improvement over the baseline network model on different datasets and compared with most other mainstream speech emotion recognition models.

To further analyze the experimental results, the confusion matrix is used to analyze the classification results in this paper, as shown in Figures 5 to 7 below.

Fig.5 Confusion matrix of DYCNN+Bi-LSTM in CASIA dataset

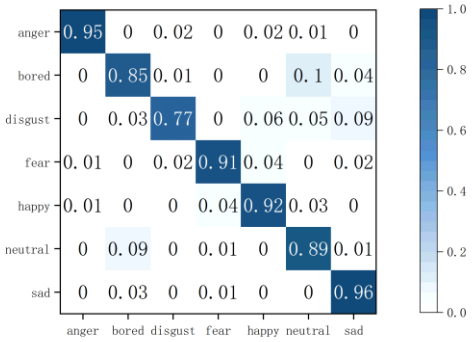

Fig.6 Confusion matrix of DYCNN+Bi-LSTM in EmoDB dataset

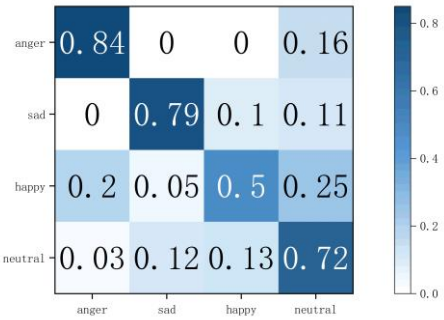

Fig.7 Confusion matrix of DYCNN+Bi-LSTM in IEMOCAP dataset

As shown in Figure 5, the recognition rates of anger and sadness in CASIA dataset are the highest, reaching 75% and 72% respectively, while the recognition rates of happiness and surprise are lower, both of which are less than half, among which happy samples are more misclassified as surprise or anger, and surprised samples are more misclassified as anger, fear and happiness; as shown in Figure 6, in EmoDB dataset The recognition rate of sadness is the highest, reaching 96%, while the recognition rates of anger, fear and happiness are all over 90%, while the recognition rate of disgust is the lowest compared with other emotions, only 77%, and is mainly misclassified as sadness or happiness; as shown in Figure 7, the recognition rate of anger in the IEMOCAP dataset is the highest, reaching 84%, and the recognition rate of sadness is also close to 80%, while the recognition rate of happiness is the lowest, only 50%, being misclassified as neutral and angry

more often. Combining the above three tables, it can be seen that emotions with high activation such as anger are more easily classified as other emotions with high activation, for example, 9% of the anger samples in Table 5 are misclassified as surprise and 8.5% are misclassified as happy; neutral emotions are closer to the center of these emotions, and other emotion samples are as easily classified as neutral, as shown in Figure 7. In conclusion, the average recognition rate of the proposed model in this paper is 59.08% in CASIA dataset, 89.28% in EmoDB dataset, and 71.25% in IEMOCAP dataset, which proves the superiority of the method in this paper.

4. Conclusion

This paper improves the classification network model of speech emotion recognition system, and proposes a speech emotion recognition network model using dynamic convolutional neural network instead of traditional convolutional neural network, and combining two-way long and short-term memory network. The proposed network uses dynamic convolution to improve the data redundancy problem of traditional convolutional neural network, and can extract the key features of different emotions more flexibly with the guarantee of computational power, and the fusion of two-way long and short-term memory network can more effectively and comprehensively use the weight coefficients and temporal information of each emotion feature for emotion recognition of the learned emotion features, which effectively improves the speech emotion recognition system It can improve the recognition rate of speech emotion recognition system.

However, the network proposed in this paper has only been experimented on three data sets, and there is still some room for improvement in recognition accuracy compared with other mature speech emotion recognition methods. Therefore, we will select more high-quality speech emotion datasets for further experiments in future research, and continue to study and optimize the network structure to ensure the computational power and speed of the network while improving the recognition accuracy.

References

[1] Akçay M B, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication, 2020, 116: 56-76.

[2] Abbaschian B J, Sierra-Sosa D, Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models[J]. Sensors, 2021, 21(4): 1249.

[3] Pandey S K, Shekhawat H S, Prasanna S R M. Deep learning techniques for speech emotion recognition: A review[C]//2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA). IEEE, 2019: 1-6.

[4] Issa D, Demirci M F, Yazici A. Speech emotion recognition with deep convolutional neural networks[J]. Biomedical Signal Processing and Control, 2020, 59: 101894.

[5] Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition[C]//Interspeech 2015. 2015.

[6] Kim J, Saurous R A. Emotion Recognition from Human Speech Using Temporal Information and Deep Learning[C]//Interspeech. 2018: 937-940.

[7] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern recognition, 2011, 44(3): 572-587.

[8] Fahad M S, Ranjan A, Yadav J, et al. A survey of speech emotion recognition in natural environment[J]. Digital Signal Processing, 2020: 102951.

[9] Roy T, Marwala T, Chakraverty S. A survey of classification techniques in speech emotion recognition[J]. Mathematical Methods in Interdisciplinary Sciences, 2020: 33-48.

[10] Reshma C V, Rajasree R. A survey on Speech Emotion Recognition[C]//2019 IEEE International Conference on Innovations in Communication, Computing and Instrumentation (ICCI). IEEE, 2019: 193-195.

[11] Ai X, Sheng V S, Fang W, et al. Ensemble learning with attention-integrated convolutional recurrent neural network for imbalanced speech emotion recognition[J]. IEEE Access, 2020, 8: 199909-199919.

[12] Hajarolasvadi N, Demirel H. 3D CNN-based speech emotion recognition using k-means clustering and spectrograms[J]. Entropy, 2019, 21(5): 479.

[13] Iqbal A, Barua K. A real-time emotion recognition from speech using gradient boosting[C]//2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2019: 1-5.

[14] Ringeval F, Eyben F, Kroupi E, et al. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data[J]. Pattern Recognition Letters, 2015, 66: 22-30.

[15] Garg U, Agarwal S, Gupta S, et al. Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma[C]//2020 12th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2020: 87-91.

[16] Eddy S R. Profile hidden Markov models[J]. Bioinformatics (Oxford, England), 1998, 14(9): 755-763.

[17] Reynolds D A. Gaussian mixture models[J]. Encyclopedia of biometrics, 2009, 741(659-663).

[18] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[19] Jain A K, Mao J, Mohiuddin K M. Artificial neural networks: A tutorial[J]. Computer, 1996, 29(3): 31-44.

[20] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1: 81-106.

[21] Peterson L E. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2): 1883.

[22] Kwon S. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach[J]. Expert Systems with Applications, 2021, 167: 114177.

[23] Kwon S. A CNN-assisted enhanced audio signal processing for speech emotion recognition[J]. Sensors, 2020, 20(1): 183.

[24] Wang J, Xue M, Culhane R, et al. Speech emotion recognition with dual-sequence LSTM architecture[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6474-6478.

[25] Chen Y, Dai X, Liu M, et al. Dynamic convolution: Attention over convolution kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11030-11039.

[26] Yang B, Bender G, Le Q V, et al. Condconv: Conditionally parameterized convolutions for efficient inference[J]. arXiv preprint arXiv:1904.04971, 2019.

[27] Zhang Y, Zhang J, Wang Q, et al. Dynet: Dynamic convolution for accelerating convolutional neural networks[J]. arXiv preprint arXiv:2004.10694, 2020.

[28] Wen H, You S, Fu Y. Cross-modal dynamic convolution for multi-modal emotion recognition[J]. Journal of Visual Communication and Image Representation, 2021: 103178.