2021 6th International Conference on Computational Intelligence and Applications (ICCIA) | 978-1-6654-3933-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIA52886.2021.00036

Speech Emotion Recognition Based on Secondary Feature Reconstruction

Zhiqiang Yuan[1,2,3,4],Shuoke Li[1,2,*],Wenkai Zhang[1,2],Runyan Du[1,2,3,4],Xian Sun[1,2],Hongqi Wang[1,2]
[1]Aerospace Information Research Institute, Chinese Academy of Sciences, China
[2]Key Laboratory of Network Information System Technology, Institute of Electronics,
Chinese Academy of Sciences, China
[3]University of Chinese Academy of Sciences, China
[4]School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, China
*Correspondence
e-mail:yuanzhiqiang19@mails.ucas.ac.cn; lisk@aircas.ac.cn; zhangwk@aircas.ac.cn;
durunyan19@mails.ucas.ac.cn; sunxian@mail.ie.ac.cn; wiecas@sina.com

***Abstract*—In the field of speech emotion recognition, most methods usually extract the audio spectrum first and then use image classification models to identify emotion categories. However, the spectrum contains timing information, so it is unreasonable to use the image classification model in natural scenes. Based on this, a Res-Trans model is proposed, which performs secondary time-series feature reconstruction on the extracted audio features. Compared with traditional methods, the Res-Trans method achieves a new state-of-the-art performance on multimodal emotion recognition dataset. At the same time, we propose a Horizontal Mixup data enhancement method suitable for audio spectrum enhancement, and experiments have verified the effectiveness of proposed method. Finally, we add an extra voiceprint recognition task to regularize the feature extraction network. After the integration of models, our model won first place in the 2020 iFLYTEK Multimodal Emotion Analysis and Recognition Challenge. The proposed Res-Trans model will be published soon[1].**

***Keywords*—speech emotion recognition, Res-Trans, Horizontal Mixup, deep learning**

I. Introduction

Language is the primary medium for human beings to express their feelings to the outside world, and human emotions directly affect their intonation. With the development of deep learning, how to make the machines automatically judge human emotions, such as happiness, sadness, anger, and calm, has gradually become an interesting and challenging research direction in speech research [1] [2]. In human-computer interaction, whether people's emotions can be acquired accurately will significantly impact the accuracy of human-computer dialogue. Therefore, Speech Emotion Recognition (SER) plays an important role in the human-computer interaction scene.

In the general speech emotion recognition system, audio features are first extracted from the original speech signal and then classified by different classifiers [3]. Mel Spectrum Coefficient (MFSC), Mel Spectrum Cepstral Coefficient (MFCC), and phoneme are the three most commonly used features in speech signal feature extraction. Wu et al. [4] compared the effects of speech duration, energy, pitch, and MFCC in audio classification. Eyben et al. [5] suggested using the Geneva minimalistic acoustic parameter set to perform downstream tasks of speech tasks. Researchers use different methods to model different types of features. For phoneme features, Schuller et al. [6] proposed a hidden Markov model which can be applied to speech emotion recognition, while Kim et al. [7] used the unsupervised machine learning algorithm KNN to perform audio MFCC features classification. After deep learning entered the field of speech recognition, speech emotion recognition technology has developed rapidly. In speech emotion recognition, Han et al. [1] is recognized as the first researcher who applied deep learning to speech emotion recognition. Lim et al. [8] have achieved good results on the IEMOCAP dataset by changing the different fusion stages of audio features. Niu et al. [9] extracted spectra of different sizes and constructed a deep convolutional neural network to classify audio features. Yenigalla et al. [10] combined phoneme embedding and spectrogram to construct a new fusion model, which achieved high accuracy in emotion recognition. However, although the previous methods have achieved good results, they still have room for improvement because of the lack of audio feature extraction and the lack of consideration of modeling the low-level features in terms of time sequence [11].

In this paper, we are inspired by the transformer structure in natural language processing [12] and build a self-attention network to perform secondary feature reconstruction, which finally achieves higher speech emotion classification accuracy. We have designed a Horizontal Mixup data enhancement method in the field of SER. This method reduces the Res-Trans model's error rate in the multimodal emotion recognition dataset by 1.3%, and dramatically improves the model's robustness. We have added the additional task of voiceprint recognition to the SER task, thus improving the model's prediction accuracy on the multimodal emotion recognition dataset. On this dataset, the Res-Trans model's accuracy

[1]https://github.com/xiaoyuan1996/Res-Trans

978-1-6654-3933-6/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCIA52886.2021.00036

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on April 21,2022 at 04:00:15 UTC from IEEE Xplore. Restrictions apply.

reaches 96.6%, and with the multimodal integration method's help, we won first place in the 2020 iFLYTEK Multimodal Emotion Analysis and Recognition Challenge.

II. Dataset

The multimodal emotion recognition (mer) dataset used in this study can be requested from the Research Center of Intelligent Acoustics and Immersion Communication of Northwestern Polytechnical University. Mer dataset contains voice, electrocardiogram (ECG), and electroencephalogram (EEG) data under different emotional arousal conditions. The challenge dataset includes the physical, psychological and behavioral data of 29 subjects under the interference of four kinds of emotions: happiness, sadness, anger, and gentleness. All data were collected in the natural environment and low-noise ideal environment. Each type of data lasts about 2 hours. The natural environment data refers to natural environment interference, including noise, reverberation, and electromagnetic interference while the noise ideal environment data were collected in Northwestern Polytechnical University's anechoic room, which greatly reduces the noise and reverberation.

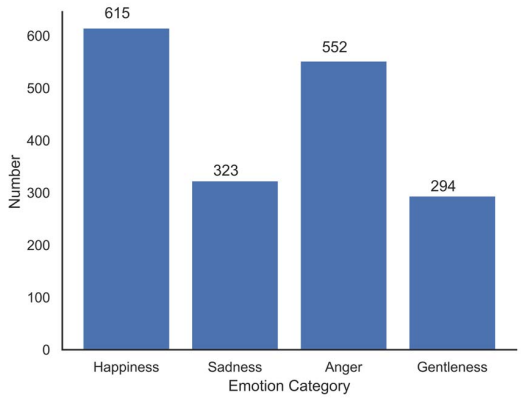


Fig. 1. Category distribution of multimodal emotion recognition dataset.

In our experiment, as shown in Fig. 1, the whole dataset is divided into four parts: happiness(34.4%), sadness(18.1%), anger(30.9%), and gentleness(16.4%), and there are 1784 training samples in total. There are 929 samples in the test set, which is more than half of the training set. Due to the lack of training samples in the mer dataset, it is more challenging than the traditional IEMOCAP [13] dataset. Simultaneously, unlike the IEMOCAP dataset, the speech data in mer dataset are all in Chinese, so this task belongs to speech emotion recognition under the Chinese scene.

III. Method

In this section, we first give an overview of audio feature extraction methods and data enhancement methods, and then we propose the Horizontal Mixup data enhancement method to improve the robustness of the model further. In terms of the model, we propose the Res-Trans speech emotion recognition

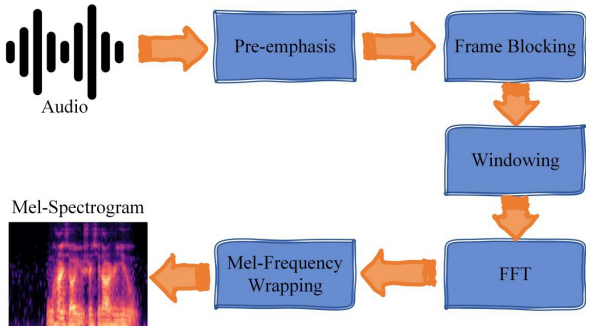


Fig. 2. Data preprocessing process.

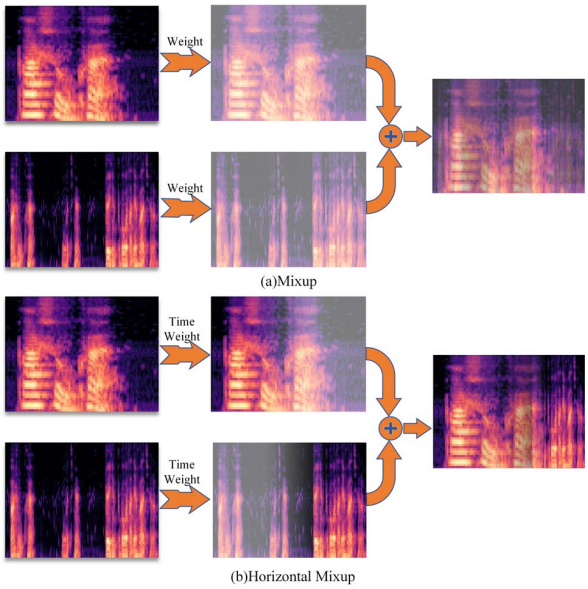


Fig. 3. Comparison of Horizontal Mixup method and Mixup method.

network. The network uses ResNet-18 with channel attention to extracte audio spectrum features and then uses the self-attention module to perform a secondary reconstruction of the spectrum features in terms of timing.

A. Extraction of Mel Spectrum

As shown in Fig. 2, for the original input audio, we use pre-emphasis to filter the signal's noise and then divide a complete audio signal into several parts. We used the Hamming window to process the framed signal and then transform the signal into the frequency domain signal by fast Fourier transform to prevent spectrum leakage. To make the signal transformation consistent with the human ear, we performed Mel spectrum conversion on the signal and then got the final spectrum image.

B. Data Enhancement

The shortage of training samples is an important factor restricting the performance of the model. Therefore, it is necessary to extend the training samples through appropriate data enhancement methods to improve the model's robustness.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on April 21,2022 at 04:00:15 UTC from IEEE Xplore. Restrictions apply.

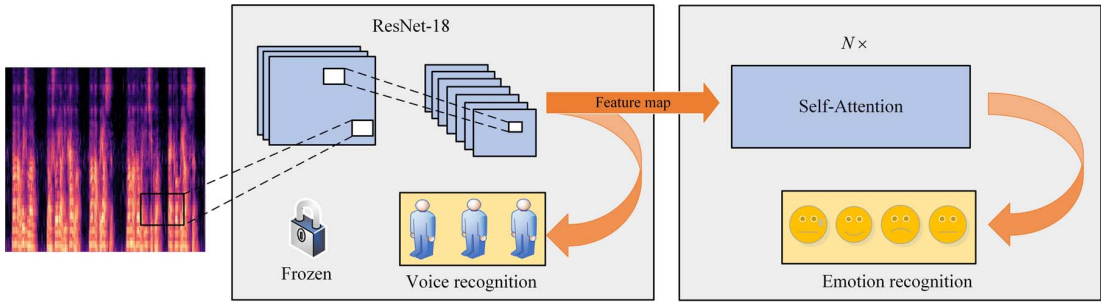


Fig. 4. Res-Trans model architecture. Our model firstly uses the voiceprint recognition task to regularize the audio feature extraction network, and then uses the self-attention module to perform a secondary reconstruction of the spectrum features in terms of timing.

After research, we found four more effective data enhancement methods in SER tasks: Cutout, Salt and pepper noise, Random cut, and Mixup. These four methods are transplanted from the classical natural image data enhancement methods. In the fourth method, the Mixup method, we have made some improvements to improve the model's performance further.

Unlike natural images, the spectrogram represents the time series on a horizontal scale, while the vertical direction represents the amplitude at different frequencies. Therefore, using the traditional Mixup method will make the spectrum of both ends overlap entirely, equivalent to the complete repetition of two speech segments in a period.

Therefore, the Mixup method is not suitable to be directly applied to audio spectrum images. For this reason, we considered the timing information and proposed the Horizontal Mixup data enhancement method. As shown in Fig. 3, we change the weights of the two spectra at the pixel level according to the time sequence, and then add the two images at the pixel level, specifically:

$$I_3(t) = min(1, w(t))I_1(t) + max(0, (1 - w(t)))I_2(t) \quad (1)$$

$$w(t) = c \times (a + b \times \frac{t}{T})(0 < a < 1, b > 0) \quad (2)$$

among them, a and b are the parameters that need to be set in order to generate weights, c is set to 1, T is the number of pixels in the horizontal direction of the image, $I_1(t)$ and $I_2(t)$ are the images before weighting, and $I_3(t)$ is the image after the Horizontal Mixup method.

From formula (1)(2), it can be seen that the setting of parameter a represents the weighted bias of the image, and the bias exists anywhere in time, so the setting of a will have a great influence on the result. However, for parameter b, it indicates the speed at which two audio clips overlap. With the change of the two audio clips' different mixing times, the category labels of the audio produced by the mixup become more and more blurred. Therefore, the uncertainty of the value of parameter b is relatively higher. When a, b, and c are 1, 0, and 0.5, the Horizontal Mixup method degenerates to the traditional Mixup method. The superimposed image has two labels, and multiple labels are used during training.

C. Res-Trans for SER

The traditional method of classifying speech emotion by using the spectrogram usually use a network model pre-trained in natural images. This method ignores the spectrogram's time relationship and only regards the spectrogram as an ordinary natural image. The motivation of Res-Trans comes from the secondary feature reconstruction of time-series signals. Spectrograms are different from traditional natural images. The horizontal axis of the spectrogram represents time-series information. Therefore, we consider using a time-series model to carry out secondary modeling of audio features. As shown in Fig. 4, we first used the convolution network to extract the audio features contained in the spectrogram. After the audio features are obtained, we used the transformer [12] to perform secondary construction of the spectral image features. To further enhance the model's robustness, we use an extra task of voiceprint recognition to constrain the convolution network.

1) ResNet with Channel Attention: We use ResNet-18 [14] as the feature extraction network. On the one hand, the network with the residual layer is less likely to be overfitting than VGGNet and other networks. On the other hand, the ResNet-18 network has fewer model parameters than the ResNet-34 network, so less training data is needed during training. To further adjust the network's channel correlation, we improved the network's learning performance by obtaining a spatial correlation learning mechanism. We followed the practice in [15], in ResNet-18, channel attention is added to the network to further selectively enhance the representation of useful audio features and suppress useless feature representations in the spectrograms. We remove the last classification layer and pooling layer in the ResNet-18 network and take the 512 feature maps output by the network as the audio features extracted by the convolution network and input them to the encoding network.

2) Voiceprint Recognition Task: To ensure that the features extracted by ResNet-18 are reasonable enough, we added the voiceprint recognition task. After the original spectrogram passes through the ResNet-18 network, We directly classify the 8×8 feature maps to predict the speaker. To avoid the influence

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on April 21,2022 at 04:00:15 UTC from IEEE Xplore. Restrictions apply.

TABLE I
ACCURACY COMPARISON OF DIFFERENT MODELS WHEN USING THE SAME DATA ENHANCEMENT METHOD.

	Hapiness Acc	Sadness Acc	Anger Acc	Gentleness Acc	F1-score
Capsule Network	0.963	0.973	0.869	0.849	0.898
Densenet-121	0.985	0.914	0.933	0.915	0.938
DPN-98	0.988	0.962	0.926	0.923	0.946
ResNext-29	0.978	0.978	0.92	0.948	0.948
ResNest-50	0.98	0.943	0.948	**0.952**	0.954
Resnet-Incep	0.979	0.959	0.938	0.885	0.942
Res-Trans	**0.994**	**0.979**	**0.957**	0.942	**0.966**

TABLE II
CONFUSION MATRIX OF RES-TRANS MODEL

Class Labels	Hapiness	Sadness	Anger	Gentleness
Hapiness	0.994	0.002	0.004	0.000
Sadness	0.012	0.979	0.000	0.008
Anger	0.017	0.010	0.957	0.015
Gentleness	0.004	0.037	0.017	0.942

caused by the network being too deep, we first perform the voice prediction task. After the voiceprint recognition prediction task is finished, we fix the CNN network's weight and only use the CNN network as the spectrogram's audio feature extractor.

3) Temporal Feature Reconstruction: We use the transformer network to reconstruct the features extracted from the ResNet-18 network. The transformer network is a powerful feature extractor. As shown in Fig. 5, the attention module firstly measures the similarity between the query and the key and then uses the similarity score to weigh the value, which can be written as:

$$a_{i,j} = f_{sim}(q_i, k_j) \tag{3}$$

$$\alpha_{i,j} = \frac{e^{a_{i,j}}}{\sum_j e^{a_{i,j}}} \tag{4}$$

$$\widehat{v}_i = \sum\nolimits_j \alpha_{i,j} v_j \tag{5}$$

where $q_i \in Q$ is the i^{th} query, $k_i \in K$ and $v_i \in V$ are the j^{th} key/value pair; f_{sim} is a function to calculate the similarity of each k_j and q_i; $\widehat{v}_i$ is the vector after attention weight.

After the attention calculation operation, each transformer network block contains a fully connected forward network and performs the same operation on each position's vectors. Meanwhile, a normalization layer is added to keep the same distribution of features, and a residual layer is added to avoid over-fitting of the network.

We input the audio feature map output by the ResNet-18 encoder into the self-attention module and obtain the final classification result using the linear layer.

IV. RESULTS

We carried out quantitative analysis and verification of our model and compared it with other SoTA models in the IEMOCAP dataset. We use the F1-Score indicator to evaluate

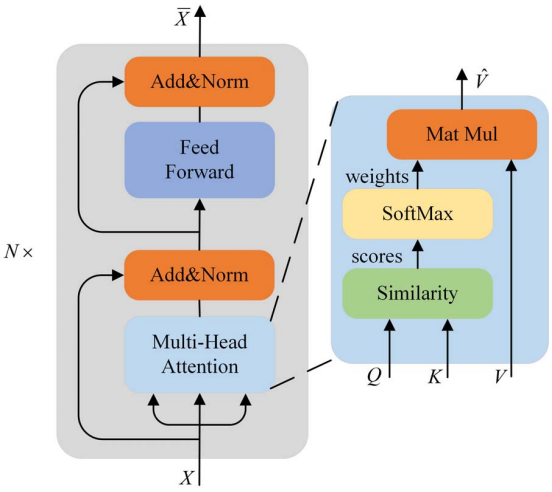


Fig. 5. Self-attention module.

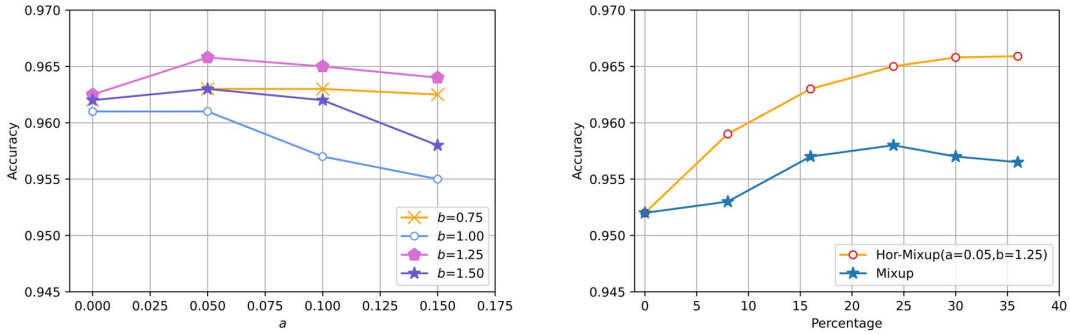
our model, and at the same time, we calculated each category's accuracy to analyze the model better.

A. Impact of Feature Reconstruction

To verify the Res-Trans model's rationality, we use the most popular models in the field of natural image and compare them with ours. Simultaneously, to maintain the experiment's consistency, we also use the inception network for secondary modeling to conclude whether it is reasonable to use the transformer model for secondary modeling of audio features.

Table I shows the experimental results on the multimodal sentiment analysis dataset. We tried several leading models in the field of natural images, such as Capsule Network, Dense Network, Dual Path Network, ResNext, and ResNest, and used these models to classify spectrograms. It can be seen from Table I that the scores of these leading classification models

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on April 21,2022 at 04:00:15 UTC from IEEE Xplore. Restrictions apply.



(a) (b)

Fig. 6. The impact of data enhancement methods on accuracy. (a)The influence of the parameters a and b in the Hor-Mixup method on the experimental results. (b)The comparison of mixup and Hor-Mixup in the case of data generated by different data enhancement methods.

TABLE III
IMPACT OF VOICEPRINT RECOGNITION.

	Hapiness Acc	Sadness Acc	Anger Acc	Gentleness Acc	F1-score
Single Task	0.989	0.967	0.950	**0.942**	0.960
Multi Task	**0.994**	**0.979**	**0.957**	0.937	**0.966**

TABLE IV
ACCURACY COMPARISON WITH SOTA MODEL ON IEMOCAP DATASET.

	Hapiness Acc	Sadness Acc	Anger Acc	Gentleness Acc	F1-score
LSTM-Based [16]	0.855	0.399	0.817	0.704	0.677
SVM-Based [17]	0.957	0.707	**0.964**	0.791	0.860
DRCNN [9]	0.966	0.891	0.899	0.850	0.904
INCEPTION-Based [18]	0.979	0.939	0.919	0.909	0.936
Res-Trans	**0.994**	**0.979**	0.957	**0.942**	**0.966**

used in the field of natural images are lower than those of our method. For the classification of Gentleness, ResNest-50 is the best, but the overall accuracy is still 1.2% lower than our method, which verifies our method's effectiveness in the task of SER.

We also conducted an experiment in which the ResNet was fixed, and the Inception network was used to perform secondary modeling of features. Compared to our Res-Trans model, the F1-score of ResNet-Inception is 2.4% lower than that of our model, which proves that the transformer model considering time series can achieve the best results in secondary feature reconstruction. Table II shows that using the confusion matrix evaluated on the Res-Trans model, F1-Score reaches 0.966.

B. Impact of Horizontal Mixup

To verify the effectiveness of the Horizontal Mixup (Hor-Mixup) data enhancement method in SER tasks, we conducted the following experiments. We use the Hor-Mixup data enhancement method to enhance the training set's data rate by 30% and observe the influence on the experimental accuracy by adjusting different values of a and b. As shown in Fig. 6(a), when the value of a is 0.05 and the value of b is 1.25, the model reaches the optimal accuracy.

Simultaneously, we have experimented that the training accuracy as the model changes with the ratio of Hor-Mixup data or Mixup data. Our experimental results are as shown in Fig. 6(b). It can be seen that as the Horizontal Mixup data increases in the training data, the training accuracy of the model also increases. Compared with the result of not using Hor-Mixup data in training, the model's accuracy is improved by 1.3%. At the same time, we compared the use of Mixup data enhancement to train the model and found that the Mixup method's contribution to the model's accuracy is weaker than that of Hor-Mixup.

C. Impact of voiceprint recognition

To evaluate the impact of the voiceprint recognition task on the experiment, we first use a single-task approach to train the Res-Trans model end-to-end and use the test set to evaluate the model. In the second set of experiments, we first used convolution networks for voiceprint recognition, and after fixing the parameters, we started to perform secondary reconstruction of audio features. The experimental data we obtained are shown in Table III. Compared to the task without speech recognition

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on April 21,2022 at 04:00:15 UTC from IEEE Xplore. Restrictions apply.

task, the network with the added task has an improvement of 0.6% in F1-score, and all categories except Gentleness have an improvement in accuracy, which undoubtedly proves the effectiveness of adding speech prediction task.

D. Comparisons with the State-of-the-art Approaches

We also compared our method with excellent models in SER tasks in recent years, and the results are shown in Table IV. Among them, for the LSTM and SVM methods, we use the methods described in the article [16] [17] to reproduce them. For DRCNN [9] and Inception-Based [18] methods, we use the data enhancement method mentioned in this paper and then perform model training. It can be seen that our model is in the optimal position in terms of accuracy, no matter for the timing-based method or the previous network such as DRCNN.

V. Conclusion

Speech emotion recognition is an essential task in human-computer interaction. If machines generate emotions in the dialogue with human beings, it will undoubtedly make the dialogue between the two sides more real and interesting. This paper proposes a new method, called Res-Trans, which firstly performs the extra voice recognition tasks and then uses a time-series model to reconstruct audio features. Besides, we propose a data enhancement method, Horizontal Mixup, which is useful for audio spectrograms. The experimental results show that this method achieves better accuracy compared with the latest technology in recent years. In the future, we will further improve this method by using model distillation.

Acknowledgment

The author is very grateful to iFlytek Co., Ltd, for providing the competition platform. Thanks to the Research Center of Intelligent Acoustics and Immersion Communication of Northwestern Polytechnical University for the dataset used in this paper.

References

[1] Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In Fifteenth annual conference of the international speech communication association.

[2] Chen, K. H., Brown, C. L., Wells, J. L., Rothwell, E. S., Otero, M. C., Levenson, R. W., & Fredrickson, B. L. (2020). Physiological linkage during shared positive and shared negative emotion. Journal of Personality and Social Psychology.

[3] Akcay, M. B., & Oguz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56-76.

[4] Wu, D., Parsons, T. D., & Narayanan, S. S. (2010). Acoustic feature analysis in speech emotion primitives estimation. In Eleventh Annual Conference of the International Speech Communication Association.

[5] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andr, E., Busso, C., ... & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE transactions on affective computing, 7(2), 190-202.

[6] Schmitt, M., Ringeval, F., & Schuller, B. W. (2016, September). At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In Interspeech (pp. 495-499).

[7] Kim, Y., & Provost, E. M. (2013, May). Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 3677-3681). IEEE.

[8] Lim, W., Jang, D., & Lee, T. (2016, December). Speech emotion recognition using convolutional and recurrent neural networks. In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA) (pp. 1-4). IEEE.

[9] Niu, Y., Zou, D., Niu, Y., He, Z., & Tan, H. (2017). A breakthrough in speech emotion recognition using deep retinal convolution neural networks. arXiv preprint arXiv:1707.09917.

[10] Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., & Vepa, J. (2018, September). Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In Interspeech (pp. 3688-3692).

[11] Kwon, S. (2020). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Systems with Applications, 114177.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

[13] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[15] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[16] Atmaja, B. T., & Akagi, M. (2019, July). Speech emotion recognition based on speech segment using lstm with attention model. In 2019 IEEE International Conference on Signals and Systems (ICSigSys) (pp. 40-44). IEEE.

[17] Prasetya, M. R., Harjoko, A., & Supriyanto, C. (2019, December). Speech Emotion Recognition of Indonesian Movie Audio Tracks based on MFCC and SVM. In 2019 International Conference on contemporary Computing and Informatics (IC3I) (pp. 22-25). IEEE.

[18] Singh, C., Kumar, A., Nagar, A., Tripathi, S., & Yenigalla, P. (2019, December). Emoception: An Inception Inspired Efficient Speech Emotion Recognition Network. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 787-791). IEEE.

Authorized licensed use limited to: University of Chinese Academy of SciencesCAS. Downloaded on April 21,2022 at 04:00:15 UTC from IEEE Xplore. Restrictions apply.