Improving Speech Emotion Recognition by Fusing Pre-trained and Acoustic Features Using Transformer and BiLSTM

Zheng Liu, Xin Kang, Fuji Ren

Improving speech emotion recognition by fusing pre-trained and acoustic features using Transformer and BiLSTM

Zheng Liu*, Xin Kang, Fuji Ren**

**Tokushima University*
School of Information Faculty of Engineering,
Tokushima
770-8506,
Japan
gentlefishliu@126.com

**Tokushima University*
School of Information Faculty of Engineering,
Tokushima
770-8506,
Japan
kang-xin@is.tokushima-u.ac.jp

*** Tokushima University*
School of Information Faculty of Engineering,
Tokushima
770-8506,
Japan
ren@is.tokushima-u.ac.jp

ABSTRACT: *With the emergence of machine learning and the deepening of human-computer interaction applications, the field of speech emotion recognition has attracted more and more attention. However, due to the high cost of speech emotion corpus construction, the speech emotion datasets are scarce. Therefore, how to obtain higher accuracy of recognition under the condition of limited corpus is one of the problems of speech emotion recognition. To solve the problem, we fused speech pre-trained features and acoustic features to enhance the generalization of speech features and proposed a novel feature fusion model based on Transformer and BiLSTM. We fused the speech pre-trained features extracted by Tera, Audio Albert, and Npc with the acoustic features of the voice, and conducted experiments on on the*

CASIA Chinese voice emotion corpus. The results showed that our method and model achieved 94% accuracy in the Tera model.

KEYWORDS: *Speech emotion recognition, Speech representation learning, feature fusion Transformer*

1. Introduction

Speech emotion recognition refers to a signal processing task that extracts emotional features from speech digital signals and recognizes the specified emotions. The research on making robots expresses emotion has attracted the attention of many researchers (Ren 09). Emotion recognition plays an important role in the application of human-computer interaction (Ren *et al.*, 2020). Speech not only contains textual information, but also contains rich emotional information. Therefore, speech emotion recognition has gradually become one of the topics that has received widespread attention in the field of speech signal processing. Limited by the high cost of constructing speech emotion corpus, it is a challenging task facing current speech emotion recognition to obtain a higher recognition rate on a limited speech emotional corpus.

In the past ten years, the emergence of machine learning has greatly promoted the development of various fields of signal processing including image, text, and speech (Liu *et al.,* 2020) (Deng *et al.,* 2020) (Huang *et al.,* 2020). For machine learning, the feature processing engineering of data is indispensable, and excellent feature construction is very important for the improvement of recognition accuracy. Data and features determine the upper limit of recognition accuracy, and machine learning models and algorithms constantly approach this upper limit. Regarding the commonly used features in the field of speech emotion recognition, after many literature surveys (Akay *et al.,* 2020) (Swain *et al.,* 2018), we found that the speech features used for speech emotion recognition usually include speech acoustic features, deep features, and hybrid features. Among them, the acoustic features of speech include traditional speech parameters such as F0, formant, signal energy, waveform MFCC, Mel cepstrum, and Fbank. Deep features refer to the features extracted from the original speech waveform or spectrum using deep learning neural network models such as CNN, RNN, DNN, or the pre-trained models. Hybrid features refer to features that are combined with language context, combined with other modal features such as facial expressions, text, and voice features for speech recognition.

Due to the high cost of constructing speech emotion data sets, the emotion data sets are scarce. Improving recognition accuracy on a small amount of data sets has always been a challenging task in the field of speech emotion recognition. At present, more and more researches are no longer satisfied with the construction of a single

emotional feature. These studies have enriched the diversity of the features of a single sample in the construction of feature engineering, but still cannot solve the problem of poor generalization of voice features caused by the rare corpus.

To solve the excessive dependence of deep learning on data, many feature extraction schemes based on transfer learning technology have appeared in recent years (Zhuang *et al.*, 2020). Transfer learning learns new knowledge using existing knowledge, and then finds the similarities between existing knowledge and new knowledge, focusing on storing existing problem-solving models, and using them for other different but related problems. Using the idea of transfer learning, a pre-trained model built on a large-scale data set can effectively improve the generalization ability of features in a small data set. Our work combines the voice transfer learning method and proposes a novel deep learning model that combines traditional acoustic features and pre-trained features and achieved excellent results in the CASIA dataset experiment.

Our work mainly has the following contributions:

1. To improve the recognition rate of speech emotion recognition, we used speech transfer learning technology for the first time, combined pre-trained features and acoustic features, and made relevant explorations to improve the generalization ability of features.

2. Based on the Transformer and BiLSTM models, we proposed a novel feature fusion model, which effectively fuses pre-trained features and acoustic features of different maximum lengths and dimensions.

3. After experiments on the CASIA dataset, the Chinese emotional speech dataset, our proposed approach achieved excellent results.

The rest of our paper is arranged as follows. The second section introduces related work on speech emotion recognition, speech pre-trained models, and the model used in our experiment. The third part describes the details of the models and methods we proposed. The fourth part shows the details of our experiment, including CASIA dataset, feature extraction, and experimental results. The last part summarizes our work and describes plans for future work.

2. Related Work

In the speech emotion recognition system, speech feature extraction and processing, as well as the construction of algorithm models, are very important for improving the ability of speech emotion recognition. In recent years, speech acoustic features and acoustic low-level feature descriptors using statistical methods have been widely used in various recognition models (Byun *et al.*, 2021). Ho et al. (Ho *et*

al., 2020) used opensmile to extract the features of LLDs, combined with RNN and attention mechanism, and achieved good recognition results. Deep features are mainly built around various spectrograms such as Mel spectrogram and MFCC spectrum of speech, combined with deep learning model learning. Yu-An Chung et al. (Kwon *et al.,* 2021) proposed the MLT-DNet model, which took the original waveform of the speech as input, and achieved a high recognition rate on the IEMOCAP and EMODB data sets. To enhance the diversity of voice features and solve the single problem of voice features, some methods related to speech feature fusion are constructed. The paper (Ho *et al.,* 2020) combines the attention mechanism and the RNN model to fuse speech features and text features to enhance the accuracy of emotion recognition.

The features used in traditional speech emotion recognition are all based on the extraction of individual speech samples. In the case of insufficient data set size, over-fitting is prone to occur, which will lead to the problem of low speech emotion recognition rate. In recent years, to improve the robustness of speech features, research on speech unsupervised representation learning has become more and more active (Yu-An *et al.,* 2018) (Steffen *et al.,* 2019) (Anonymous *et al.,* 2020) (Jan *et al.,* 2019). The main motivation of this research is to extract higher-level feature expressions of speech. The features of speech itself are learned in a large-scale data set, and the trained representation model can be used for many downstream tasks.

The speech features extracted through representation learning enhance the generalization of traditional features. In this paper, we use voice pre-trained features and fuse individual acoustic features to improve the performance of the neural network model. The model feature fusion is roughly divided into two categories, pre-fusion and post-fusion. The papers (Wang *et al.,* 2020) (Zhang *et al.,* 2021) only used simple feature splicing after extracting features of different types and dimensions of speech, which is insufficient to express the corresponding relationship between the features. The emergence of transformer (Vaswani *et al.,* 2017) can focus on the correspondence of local information between sentences, so it is more efficient for different types of feature fusion.

This paper proposes a novel feature fusion model based on transformer and BiLSTM model. The model we proposed can effectively fuse the pre-trained features of different maximum lengths and feature dimensions with traditional acoustic features, which greatly improves the prediction accuracy of emotion recognition.

3. Method

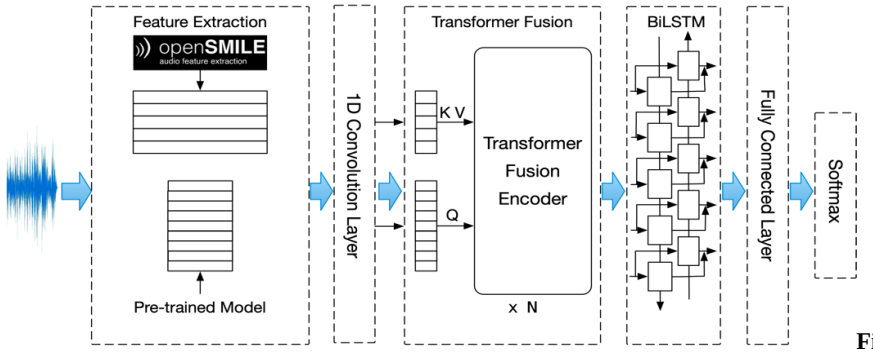

Figure 1. *Overall architecture for fusing pre-trained features and acoustic features*

In this section, we will introduce our new framework for speech emotion recognition. Our framework combines the acoustic features of traditional speech with pre-trained features. After the proposed feature fusion model, the two features are fully fused, and finally achieve the purpose of improving the accuracy of speech emotion recognition. As shown in the figure 1, the entire framework mainly includes a feature construction part, a 1D convolution module, a Transformer-based feature fusion module, a BiLSTM modle, the final fully connected layer and Softmax module.

In the feature extraction part, we used the OpenSmile tool to split each utterance into segments with a length of 200ms. Each sub-segment was extracted to features of 1583 dimensions, and a total of n*1583 feature blocks were constructed, where n is the number of speech segments. In addition, to make up for the limitations of traditional features, we fused the traditional acoustic features and pre-trained features extracted from the latest speech representation learning models NPC, Audio Albert, Tera, which are trained in the large-scale corpus.

The feature fusion part will first undergo 1D convolution processing, and the two features will be unified into vectors of different lengths but the same dimension, and then sent to the Transformer attention fusion model for further fusion. As shown in the figure 2, Transformer abandons the traditional CNN and RNN structure. The entire network structure is entirely composed of Attention mechanism, which increases the training speed and can effectively capture the relationship between the input units. In our experiment, we used a 6-layer Transformer encoder to fuse the pre-trained and acoustic features. The attention mechanism for a sentence in the traditional Transformer is represented by formula 1, where Q represents the query vector, and KV represents the vector being queried.

$$\text{Attention}\,(Q,K,V)=\text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

In our method, since speech pre-trained features and traditional acoustic features have different maximum lengths and dimensions, we first use 1D convolutional network to convert the acoustic features and pre-trained features to the same dimensional features, then feed the vectors to the fusion model. The following formula 2 is used to fuse the features. X_α and X_β respectively represent the acoustic features and pre-trained features. We define the Querys as $Q_\alpha=X_\alpha W_{Q\alpha}$, Keys as $K_\beta=X_\beta W_{K\beta}$ and Values as $V_\beta=X_\beta W_{V\beta}$. The adaptation from acoustic features to pre-trained features is presented as $PF_{\beta\to\alpha}(X_\alpha, X_\beta)$.

$$\begin{aligned} PF_{\alpha\to\beta}(X_\alpha,X_\beta) &= \text{Softmax}\left(\frac{Q_\beta K_\alpha^{\mathrm{T}}}{\sqrt{d_k}}\right)V_\alpha \\ &= \text{Softmax}\left(\frac{X_\beta W_{Q_\beta} W_{K_\alpha}^{\mathrm{T}} X_\alpha^{\mathrm{T}}}{\sqrt{d_k}}\right)X_\alpha W_{V_\alpha} \end{aligned} \tag{2}$$

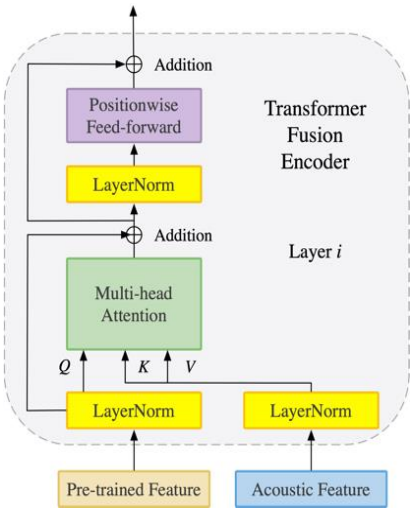

Figure 2. *Transformer fusion module of pre-trained features and traditional acoustic features*

After the feature fusion of the Transformer mechanism, to further enhance the spatial timing relationship of the hidden features, we send the output hidden vectors of the Transformer to the BiLSTM, as shown in the figure 3. The LSTM model

solves the problem of gradient disappearance and gradient explosion caused by the long-time sequence segment in the back propagation process of the traditional RNN model. We use the BiLSTM to further process the features outputted by the Transformer and finally use softmax for emotional classification.

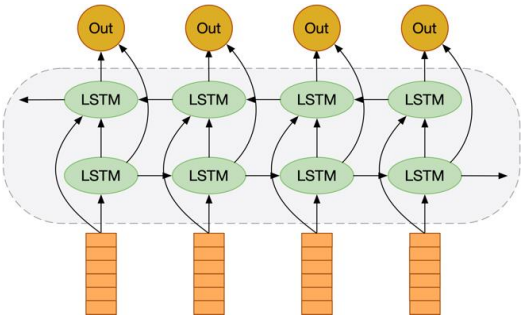

Figure 3. *Struction of BiLSTM model*

4. Experiments and Discussion

To verify the effectiveness of our proposed method, we fused the pre-trained features and traditional acoustic features using three pre-trained models respectively. We conducted experiments on the CASIA dataset, and the results showed that our proposed method framework achieved excellent results.

4.1. *CASIA dataset*

The CASIA dataset was recorded by the Institute of Automation, Chinese Academy of Sciences. A total of 2 males and 2 females participated in the recoding of this dataset. In a pure recording environment with a signal-to-noise ratio of about 35db, based on 5 different emotions, happy, sad, angry, frightened, and neutral, it was obtained from a performance of 500 sentences of text. For voice quality, a 16kHz sampling rate and 16bit quantization standard are used. Finally, after listening and screening, a total of 9600 utterances were retained. In our experiment, 6000 utterances were used and divided into training set, validation set, and test set of 5440, 280, and 280 respectively.

4.2. *Feature extraction*

In the experiment, we used three speech representation models, Tera, Audio Albert, and NPC, combined with traditional acoustic features to improve the accuracy of speech emotion recognition.

TERA is a self-supervised speech pre-training model, its full name is Transformer Encoder Representations from Alteration, which is used for pre-training on many unlabelled speech to obtain the Transformer encoding model by masking the speech spectrum along three orthogonal axes (Liu *et al.*, 2021). AUDIO ALBERT, also called AALBERT, uses the ALBERT self-supervised learning model, which is trained on a large-scale speech dataset, and can be used for feature extraction of downstream tasks such as speech-related tasks, or as a fine-tuning participation model training (Chi *et al.*, 2021). The full name of NPC is Non-Autoregressive Predictive Coding, which is also a self-supervised learning method. It only relies on the local information of the voice to represent the voice in a non-autoregressive manner. It has achieved good results in the voice speaker classification experiment (Liu *et al.*, 2020).

In addition, we split each speech utterance into segments with the 200ms length and then use the opensmile feature extraction tool (Eyben *et al.*, 2010) to extract for 1582 dimensional acoustic features. Then we fuse pre-trained representations and traditional speech features to enhance the generalization ability of speech features by using the proposed model.

4.3. *Evaluation metrics*

In this experiment, for the results of speech emotion recognition, we used two evaluation metrics, namely F1 and ACC. The F1 is the weighted average of precision and recall, and the F1 formula is expressed as formula 3. Among them, TP (True Positive) indicates that the model prediction result is positive, and the sample is also positive, TN (True Negative) indicates that the model prediction result is positive, but the sample is negative, FP (False Positive) indicates that the prediction is negative, and the sample is positive, FN (False Negative) indicates that the prediction is negative, and the sample is also negative.

$$F1 = \frac{2 * \text{ Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

Ac is the classification accuracy score, which refers to the percentage of all classifications that are correct. The formula is as follows:

$$\text{Ac } = \frac{TP + TF}{TP + TF + FP + FN} \tag{4}$$

4.4. *Result and analysis*

Table 1 shows the performance results of our proposed model with different pre-trained models. The performance is the worst when the acoustic features are used alone. After the pre-trained features are added, a better recognition rate is obtained. Among them, the fusion of Tera pre-trained features and acoustic features achieved the best results with F1 value of 0.942 and Ac value of 0.943.

Dataset	CASIA	
Metric	F1	Ac
Acoustic Only	0.789	0.789
Acoustic+Npc	0.861	0.861
Acoustic+Audio Albert	0.915	0.914
Acoustic+Tera	**0.942**	**0.943**

Table 1. *Performance of different pre-trained features and acoustic features on CASIA*

Figure 4 shows the detailed results of different emotion recognition of our model on the CASIA dataset. Among them, a high recognition rate of 100% was obtained on Angry, and a recognition rate of 82.14% was obtained on Happy. Some samples of Happy were incorrectly recognized as Angry.

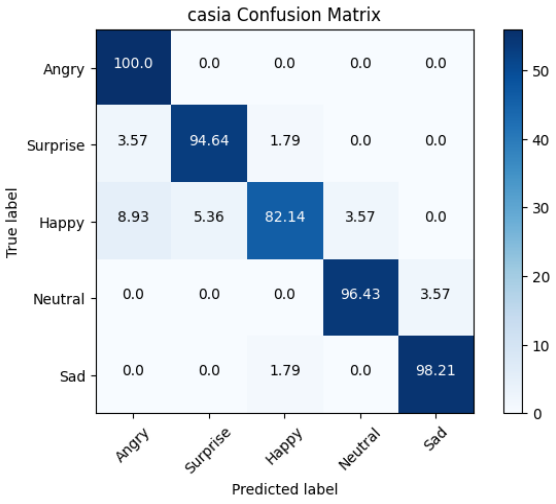

Figure 4. *Detailed results of using Tera pre-trained features and acoustic features on CASIA*

Figure 5 is the loss value change curve of different feature combinations during the training process. Due to the singularity of acoustic features, the fitting effect of the model is not good. The model combined with pre-trained features enhances the generalization ability of data features, making the model better fit the data. Compared with other loss curves, the fusion of Tera pre-trained features and acoustic features achieved the best results. It can better improve the generalization ability of features, and make the model achieve a better recognition rate.

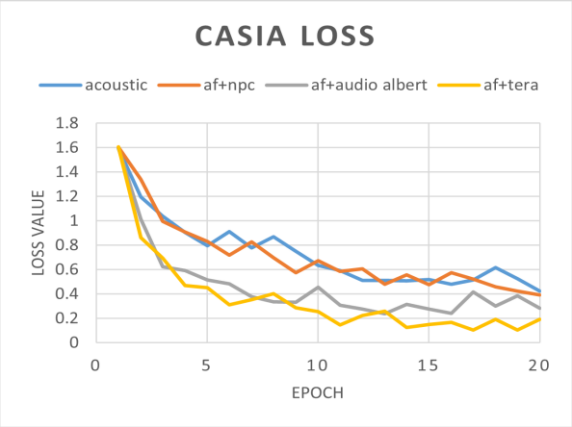

Figure 5. *Loss value of different feature combinations*

5. Conclusions

To solve the problem of low recognition accuracy caused by insufficient generalization ability of speech features in the field of speech emotion recognition, we proposed a novel feature fusion model based on Transformer and BiLSTM, which can effectively fuse speech pre-trained features and acoustic features of different lengths and dimensions .

We utilized Tera, Audio Albert, and NPC three pre-trained models and conducted experiments on CASIA dataset. The experimental results show that all the combinenation features between pre-trained features and acoustic features achieved better results. Especially the combineation between Tera pre-trained features and acoustic features, achieved a prediction accuracy of 94%.

In the future, we will further explore the application of transfer learning in the field of speech emotion recognition, try more feature fusion structures, enhance the generalization ability of speech feature expression, and try the application of speech representation learning in multi-modal emotion recognition to improve speech emotion recognition accuracy.

6. Acknowledgments

This research has been supported by JSPS KAKENHI Grant Number 19K20345 and Grant Number 19H04215.

7. References

Ren, Fuji. ”Affective information processing and recognizing human emotion.” Electronic notes in theoretical computer science 225 (2009): 39-50.

Ren, Fuji, and Yanwei Bao. ”A review on human-computer interaction and intelligent robots.” International Journal of Information Technology & Decision Making 19.01 (2020): 5-47.

Liu,Zheng, et al.”Vowel priority lip matching scheme and similarity evaluation model based on humanoid robot Ren-Xin.” Journal of Ambient Intelligence and Humanized Computing (2020): 1-12.

Deng, Jiawen, and Fuji Ren. ”Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning.” IEEE Transactions on Affective Computing (2020).

Huang,Zhong, et al.”Facial Expression Imitation Method for Humanoid Robot Based on Smooth-Constraint Reversed Mechanical Model (SRMM).” IEEE Transactions on Human-Machine Systems 50.6 (2020): 538-549.

Akay, Mehmet Berkehan, and Kaya Ouz. ”Speech emotion recog- nition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers.” Speech Communica- tion 116 (2020): 56-76.

Swain, Monorama, Aurobinda Routray, and Prithviraj Kabisatpa- thy. ”Databases, features and classifiers for speech emotion recognition: a review.” International Journal of Speech Technology 21.1 (2018): 93-120.

Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." Proceedings of the IEEE 109.1 (2020): 43-76.

Byun, Sung-Woo, and Seok-Pil Lee. "A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms." Applied Sciences 11.4 (2021): 1890.

Ho, Ngoc-Huynh, et al. "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network." IEEE Access 8 (2020): 61672-61686

Kwon, Soonil. "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach." Expert Systems with Applications 167 (2021): 114177.

Ho, Ngoc-Huynh, et al. "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network." IEEE Access 8 (2020): 61672-61686.

Yu-An Chung and James Glass, “Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech,” Interspeech 2018, Sep 2018.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, “wav2vec: Unsupervised pre-training for speech recognition,” Interspeech, Sep 2019.

Anonymous authors, “vq-wav2vec: Self-supervised learning of discrete speech representations,” in ICLR 2020 Conference Blind Submission, 2020.

Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aaron van den Oord, “Unsupervised speech representation learning using wavenet autoencoders,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2041–2053, Dec 2019.

Wang, Wei, et al. "Significance of phonological features in speech emotion recognition." International Journal of Speech Technology 23.3 (2020): 633-642.

Zhang, Shiqing, et al. "Learning deep multimodal affective features for spontaneous speech emotion recognition." Speech Communication 127 (2021): 73-81.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Liu, Andy T., Shang-Wen Li, and Hung-yi Lee. "Tera: Self-supervised learning of transformer encoder representation for speech." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 2351-2366.

Chi, Po-Han, et al. "Audio albert: A lite bert for self-supervised learning of audio representation." 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021.

Liu, Alexander H., Yu-An Chung, and James Glass. "Non-autoregressive predictive coding for learning speech representations from local dependencies." arXiv preprint arXiv:2011.00406 (2020).

Eyben, Florian, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor." Proceedings of the 18th ACM international conference on Multimedia. 2010.