Transformer encoder-based multilevel representations with fusion feature input for speech emotion recognition

He Zhengran[1] Shen Qifan[1] Wu Jiaxin[2] Xu Mengyao[3] Zhao Li[1]

([1]School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([2]School of Electronic Science and Engineering, Southeast University, Nanjing 210096, China)
([3]School of Computer Science and Software Engineering, University of Stirling, Stirling FK9 4LA, UK)

Abstract: To improve the accuracy of speech emotion recognition (SER), the possibility of applying transformer-based SER is explored. The log Mel-scale spectrogram and its first-order differential feature are fused as the input to extract hierarchical speech representations using the transformer. The effects of the variation in the number of attention heads and the number of transformer-encoder layers on the recognition accuracy are discussed. The results show that the accuracy of the proposed model increased by 13.98%, 8.14%, 24.34%, 8.16%, and 20.9% compared with that of the transformer with the Mel-frequency cepstral coefficient as featured on the ABC, CASIA, DES, EMODB, and IEMOCAP databases, respectively. Compared with recurrent neural networks, convolutional neural networks, transformer-based models, and other models, the proposed model performs better.

Key words: speech emotion recognition; transformer; multihead attention mechanism; fusion feature

DOI: 10.3969/j. issn. 1003 − 7985. 2023. 01. 008

Language is the most common tool used by human beings to communicate and express ideas in daily life. As a medium of language transmission, a speech signal contains abundant emotional information, which can reflect the psychological state of the speaker. Correspondingly, most people can automatically and effectively perceive features in other individuals' speech signals and recognize emotional information from them, which is a natural and unconscious process but a rather challenging task for machines.

Speech emotion recognition (SER) technology can bring great convenience to medical, education, and other industries. The continuous development of artificial intelligence and the in-depth research on emotion recognition will bring new breakthroughs in the field of human-computer interaction. Therefore, the study of SER has important theoretical value and significance.

In recent years, traditional networks, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), have been greatly explored in SER. In common practice, Mel-scale spectrogram, Mel-frequency cepstral coefficients (MFCCs), or other audio characters will be used as input features for SER. Most of these features are two-dimensional, making it natural to use the CNN model to process these image-like features. Furthermore, these features are extracted from speech frames, which exert a temporal meaning on them that can be treated like a sequence by an RNN[1]. Issa et al. [2] performed a one-dimensional CNN on the combination of five different audio features, achieving a high recognition rate of 86.1% on the EMODB dataset. Ref. [2] also mentioned that the use of an additional long short-term memory (LSTM) layer may lead to a good performance. As Chen et al. [3] did, a three-dimensional convolutional RNN was employed to reduce the emotionally irrelevant factors, which also shows superiority in terms of the unweighted average recall.

Although using these traditional strategies can achieve a pretty good performance on recognition and has become a consensus, researchers still attempted to use novel approaches in studies of speech signals. As such, the attention mechanism is the mainstream method[4]. Since the proposal of the transformer model by Google[5], it was received considerable attention from all fields. In the beginning, they were used in the field of natural language processing and then gradually applied to computer vision tasks.

In this paper, we propose the speech-emotion-transformer (SET) model, which uses the fusion feature as an input. We compare the accuracy of using MFCC, Mel-scale, and fusion features with this model in different datasets. Then, we compare the average accuracy among the different settings of attention heads and transformer-encoder layers. The comparison of the performance between the SET model and the traditional CNN model is also mentioned in this paper.

1 Proposed Method

In this section, we will introduce the preprocessing

Received 2022-08-15, **Revised** 2023-01-30.

Biographies: He Zhengran (1993—), male, Ph. D. candidate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@ seu. deu. cn.

Foundation item: The Key Research and Development Program of Jiangsu Province (No. BE2022059-3).

Citation: He Zhengran, Shen Qifan, Wu Jiaxin, et al. Transformer encoder-based multilevel representations with fusion feature input for speech emotion recognition[J]. Journal of Southeast University (English Edition), 2023, 39(1): 68 − 73. DOI: 10. 3969/j. issn. 1003 − 7985. 2023. 01. 008.

steps before model training and the fusion feature method first. Then, our SET model will be presented in detail.

1.1 Fusion feature

Before extracting features, we normalized each speech signal to a zero mean and unit variance so that the amplitude will have the same distribution range. Then, each speech was split into several frames of a 25 ms time span with Hanning windows, and a 10 ms time span was also applied. Next, we calculated the log Mel-scale spectrogram and limited the frame count to 300, which makes the size of the model inputs consistent. In detail, the number of Mel-filter banks was 80.

In natural language processing, for any language, the position of words in a sentence and the order in which they are arranged are very important, not only as part of the grammatical structure of a sentence but also as an important concept for expressing semantics. If a word is placed or arranged in a different order in a sentence, the meaning of the whole sentence may deviate. Because the transformer discards the CNN or RNN structure, positional encoding is necessary. Hence, to highlight the variation between speech frames in the emotion recognition task, we did not focus on semantics. Instead, we are concerned about the variation of intonation. Directly adding positional encoding to the input may introduce perturbations to the input information. Hence, we calculated the first-order differential of the extracted Mel-scale spectrogram and concatenated them together. In this way, we can finally obtain a tensor of size [160, 300] as a model input. The formula for calculating the differential feature $\delta_i(i = 1, 2, \ldots, 300)$ is given by

$$\delta_i = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2} \tag{1}$$

where $m_i(i = 1, 2, \ldots, 300)$ denotes the Mel-spectrogram vector of the i-th frame and N is the differential width.

1.2 Model structure

As shown in Fig. 1, SET contains two main parts: transformer-encoder layers and a CNN model. The encoders are used to extract different levels of feature representations from the input. In addition, the CNN module may downsample high-level representations produced by encoders and correctly distinguishes different emotions.

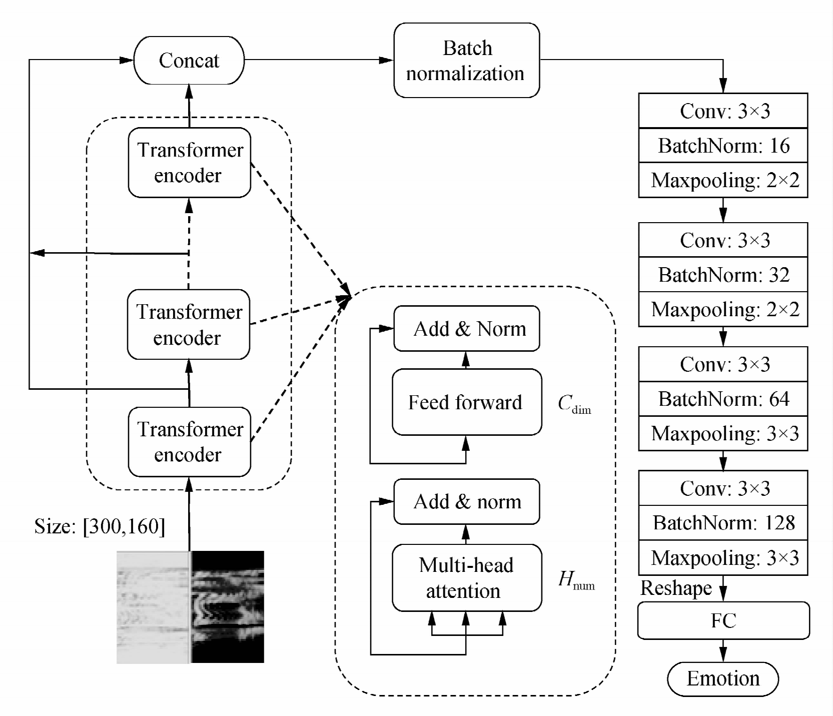

Fig. 1 SET model

1.2.1 Transformer encoders

The transformer architecture encodes an input sequence's context vectors as a set of key-value pairs with the same dimension as the input sequence length. The keys (inputs) and values (inputs' hidden states) comprise the encoder's hidden states. The output predicted at the previous timestep by the decoder is computed into a "query", and the next term in the decoder's output sequence is a mapping from the key-value pairs plus query[6]. Each output term of the decoder is a weighted sum of all values from the key-value pair's encoded representation of the input. Similar to a regular attention mechanism that decodes a weighted sum of hidden states, self-attention assigns the (alignment) weights to each value (hidden state) as a sequence-length-scaled dot product of the query with all the keys. That is, the weighted sum of all the inputs' hidden states is computed by the previous (last) term in the output sequence and the entire input sequence. This is where the global attention ability of the transformer originates[4].

The equation behind the self-attention mechanism is given by

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{n}}\right)\boldsymbol{V} \tag{2}$$

The attention can be computed by a set of queries, keys, and values packed into the matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. The scaled dot product is obtained by just scaling the dimension n of source hidden states for the sequence output at the timestep T.

The scaled dot-product self-attention ($\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$) is computed over multiple "representation subspaces". Therefore, each query, key, and value have its own weight matrix. In this way, multihead (multilayer) self-attention can compute a term in the output sequence weighted differently according to a region (subspace) of the input sequence. Each attention head in multihead self-attention still computes a scaled dot product over the entire ($\boldsymbol{K}, \boldsymbol{V}$) encoded input, just weighted differently to the input values.

The output of all the attention heads is concatenated and multiplied with a weight matrix that puts the dimension of the encoded state back to that of a single attention head. Then, a single feedforward layer can operate on the encoded latent space regardless of the number of attention heads, and a softmax prediction is computed from a weighted sum of all layers in the multihead attention architecture. The multihead attention layers are the meat of the transformer.

For the same d_{model}-dimensional queries, keys, and values, multiple learnable linear projections are performed to linearly project queries, keys, and values to the d_k, d_k, and d_v dimensions, respectively. The equation behind the multihead self-attention mechanism is

$$\begin{gathered}\text{Multihead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{concat}(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_H)\boldsymbol{W}^{\mathrm{O}} \\ \boldsymbol{h}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V)\end{gathered} \tag{3}$$

where the projections $\boldsymbol{W}_i^Q \in \mathbf{R}^{d_{\text{model}} \times d_k}$, $\boldsymbol{W}_i^K \in \mathbf{R}^{d_{\text{model}} \times d_k}$, $\boldsymbol{W}_i^V \in \mathbf{R}^{d_{\text{model}} \times d_v}$, and $\boldsymbol{W}_i^O \in \mathbf{R}^{(Hd_v) \times d_{\text{model}}}$ are the parameter matrices to be learned, and H is the number of heads.

In this study, we implemented a multilayer transformer-encoder framework. Then, we extracted and concatenated the outputs of all encoder layers. As shown in Fig. 1, we denote the number of adopted transformer-encoder layers as L_{num}, which obtains multilevel emotion representations with L_{num} feature maps. Meanwhile, we denote the feedforward dimension as C_{dim} and the number of attention heads as H_{num}. H_{num} and L_{num} are two important factors affecting the final recognition effect. The impact of their various values will be discussed in the experimental section. Finally, C_{dim} was fixed to 1 024, whereas 768 is also appropriate.

1.2.2 Convolution module

Given a multilevel emotional representation output by transformer encoders, the CNN module was used to downsample the representations and learn the spatial features from them. In this study, the CNN block contains four similar convolutional layers. From the first convolutional layer to the last, the number of acquired feature maps increased from 16 to 128 exponentially by 2. The size of the convolution kernel of each layer was fixed to 3 ×3. After each convolution layer, implementing a batch normalization operation can improve the recognition accuracy. To achieve downsampling, a max-pooling layer was added, and the first two pooling sizes were both 2 × 2, whereas the latter two were 3 ×3. Finally, the output resulting from the CNN module was flattened into a one-dimensional vector and then passed through a fully connected layer.

2 Experiments

To evaluate the performance of the transformer on SER tasks, we performed SER experiments on five databases, namely, ABC, CASIA[7], DES[8], EMODB[9], and IEMOCAP[10].

We split all the datasets into a training set and test set at a ratio of 7∶3. To ensure that the data for each emotion can be balanced-distributed between the training set and test set, we divided each emotional data into the same proportion instead of splitting randomly throughout the entire dataset. Then, each audio was divided into 3 s segments.

The SET module was implemented with the PyTorch toolkit. The model was optimized by minimizing the cross-entropy loss function with a mini-batch of four parallel samples using the Adam optimizer. The learning rate dropped from the initial 10-4 to 10-5, and the epoch number was set to 50. The remaining parameters were set as default values.

2.1 Feature comparison

As mentioned above, instead of using MFCCs, Mel-scale spectrogram, or other audio characters, we fused a log Mel-scale spectrogram and its first-order differential feature as the input of SET.

We compared the accuracy of using the MFCC, Mel-scale, and fusion features with SET in different datasets. We fixed the number of attention heads as 8 and the number of transformer-encoder layers as 8. The results are shown in Fig. 2.

Evidently, using the fusion feature achieves a good performance.

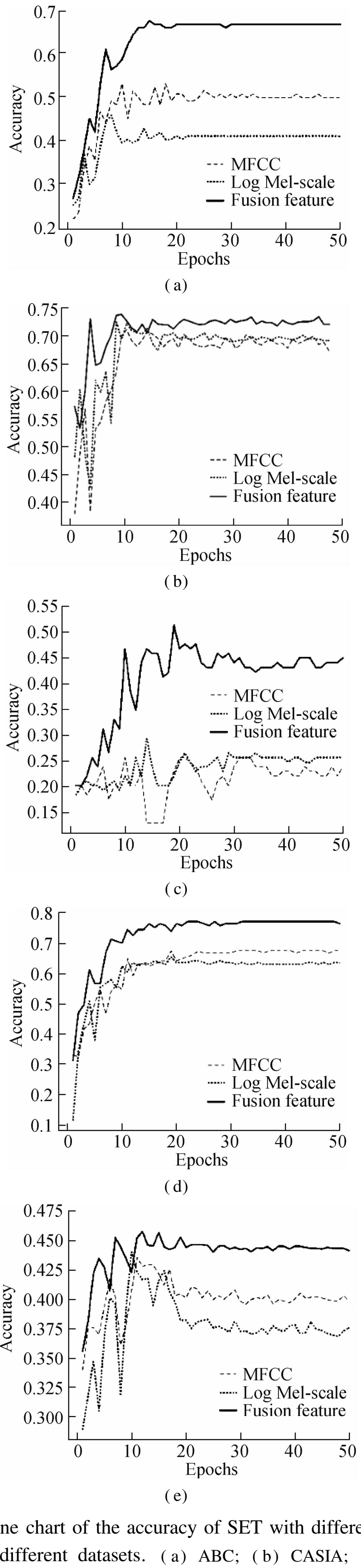

Fig. 2 Line chart of the accuracy of SET with different feature inputs on different datasets. (a) ABC; (b) CASIA; (c) DES; (d) EMODB; (e) IEMOCAP

2.2 Key hyperparameter comparison

As mentioned before, our proposed SET model has two vital hyperparameters: the number of attention heads H_{num} and the number of transformer-encoder layers L_{num}. In this section, we will discuss the influence of the variation of the two parameters on recognition accuracy. As shown in Tab. 1, when we used a combination of parameters of eight-layer transformer encoders and eight heads of multihead attention, we achieved the best performance on the ABC and DES datasets. Meanwhile, when we used a combination of parameters of 10-layer transformer encoders and 10 heads of multihead attention, the model achieved the best performance on the CASIA and EMODB datasets.

Tab. 1 Accuracy of various settings of L_{num} and H_{num} %

L_{num}	H_{num}	ABC	CASIA	DES	EMODB	IEMOCAP
4	4	68.00	61.00	39.45	64.97	62.71
6	4	63.20	76.83	42.20	71.97	63.83
6	8	60.00	71.36	47.71	73.25	64.57
8	8	66.40	73.31	50.46	68.15	64.42
8	10	66.40	76.30	44.40	72.80	62.88
10	10	64.80	79.39	46.79	73.89	63.97
10	12	60.20	71.36	42.40	70.20	63.24
12	12	62.40	72.33	44.66	66.83	62.20

Meanwhile, with the combination of the parameters of 6-layer transformer encoders and eight heads of multihead attention, the model achieved the best performance on the IEMOCAP database.

The result also reveals that using more encoders and heads can get a better overall performance, but quite large values may also cause the opposite effect. Eventually, we obtained the best SET models on two datasets, and we will show their recognition effect in multiple experiments in the next section.

2.3 Experimental results

We compared our method with the other methods using traditional networks. We recorded the accuracy using MFCCs as features while using the RNN, CNN, and transformer and the accuracy while using SET. The comparison results are presented in Tab. 2.

Tab. 2 SER performance comparison among the RNN, CNN, transformer, and SET %

Method	RNN	CNN	Transformer	SET
ABC	32.56	41.86	52.42	66.40
CASIA	53.80	61.48	71.25	79.39
DES	30.36	25.89	26.12	50.46
EMODB	44.72	46.58	65.73	73.89
IEMOCAP	25.94	34.17	43.67	64.57

Tab. 3 presents the performance of our approach to emotion recognition compared with other methods using fusion features. The baseline model is based on fully convolutional networks. Attention-based convolutional recurrent neural network (ACRNN) cascade CNNs with LSTM applies attention mechanisms on the LSTM output. We also included the model recognition rate when using positional coding and not using positional coding.

Tab. 3 SER performance comparison among the baseline, ACRNN, and SET %

Method	Baseline	ACRNN	SET	
			With positional coding	Without positional coding
ABC	60.34	63.25	60.54	66.40
CASIA	56.24	61.20	72.32	79.39
DES	28.58	30.66	45.26	50.46
EMODB	67.68	54.88	70.64	73.89
IEMOCAP	58.38	53.84	63.24	64.57

As shown in Tab. 3, the recognition rate is lower when using unlearnable sinusoidal-based positional coding compared to not using positional coding. This fixed-position coding approach only provides absolute position relationships between frames and does not reflect changes, such as pitch and formants, from frame to frame.

3 Discussion

To investigate the role that the transformer encoder plays in the proposed model, we extracted the output of each transformer-encoder layer when using the trained model for the identification of a random sample. The outputs of the encoders constitute a multilevel representation of sentiments. Fig. 3 shows the input feature of the *fear* sample “03a04Ad.wav” in the EMODB database and the output representation after passing through each encoder.

Interestingly, the output of the encoder basically maintains a rough outline of the input feature. In other words, the output of each encoder is still a spectrogram-like representation. Although the output representation of each encoder layer is more abstract and ambiguous than the previous one, the lateral grain of the input features is still preserved. We attribute this to the encoders increasingly focusing on the parts that have the potential to contain sentiment information.

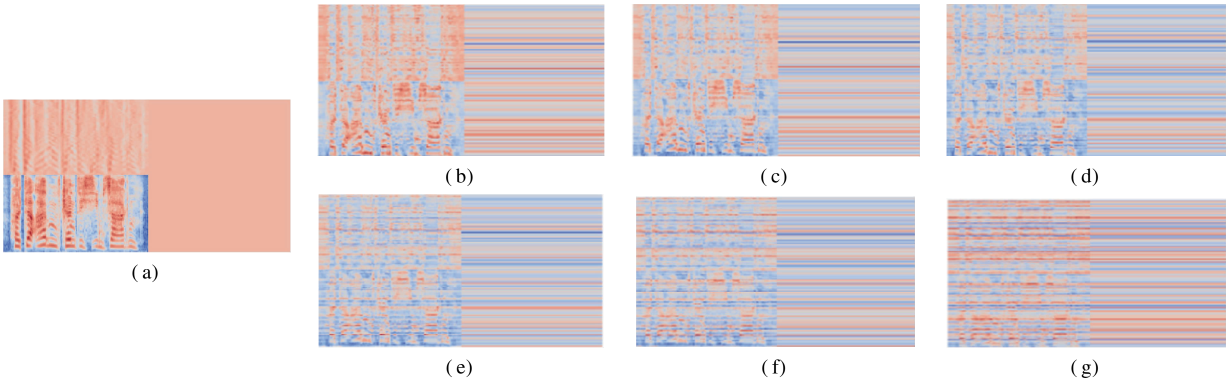

Fig. 3 Input features and multilevel emotion representation of the proposed SET model. (a) Input representation; (b) Multilevel representation 1; (c) Multilevel representation 2; (d) Multilevel representation 3; (e) Multilevel representation 4; (f) Multilevel representation 5; (g) Multilevel representation 6

4 Conclusions

1) In this study, the possibility of applying a transformer to SER is proven. SET was built with three steps: First, we fused the log Mel-scale spectrogram and its first-order differential feature as the input. Next, transformer-encoder layers were applied to extract multilevel feature representations. Finally, a common CNN module enabled the emotions to be correctly distinguished.

2) Based on the experimental results, the transformer with MFCCs as features achieved better performance than the RNN and CNN. The accuracy of the SET model increased by 13.98%, 8.14%, 24.34%, 8.16%, and 20.9% compared with the transformer with MFCCs as features on the ABC, CASIA, DES, EMODB, and IEMOCAP databases, respectively.

3) Positional coding is not necessary for SET.

References

[1] Li D, Liu J, Yang Z, et al. Speech emotion recognition using recurrent neural networks with directional self-attention[J]. *Expert Systems with Applications*, 2021, **173**(3): 114683. DOI: 10.1016/j.eswa.2021.114683.

[2] Issa D, Demirci M F, Yazici A. Speech emotion recognition with deep convolutional neural networks[J]. *Biomedical Signal Processing and Control*, 2020, **59**: 101894. DOI: 10.1016/j.bspc.2020.101894.

[3] Chen M, He X, Jing Y, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition [J]. *IEEE Signal Processing Letters*, 2018, **25**(10): 1440 – 1444. DOI: 10.1109/LSP.2018.2860246.

[4] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//2015 *International Conference on Learning Representations.* San Diego, CA, USA, 2015: 1 – 15. DOI: 10.48550/arXiv.1409.0473.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//31*st Conference on Neural Information Processing Systems*. Long Beach, CA, USA, 2017: 1－15. DOI: 10.48550/arXiv.1706.03762.

[6] Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading[C]//2016 *Conference on Empirical Methods in Natural Language Processing*. Austin, TX, USA, 2016: 16. DOI: 10.18653/v1/D16-1053.

[7] Wang K, An N, Li B N, et al. Speech emotion recognition using fourier parameters[J]. *IEEE Transactions on Affective Computing*, 2017, **6**(1):69－75. DOI: 10.1109/TAFFC.2015.2392101.

[8] Inger S E, Anya V H. *Documentation of the danish emotional speech database*(*DES*)[R]. Aalborg, Denmark: Center for Person Kommunikation, 1996.

[9] Burkhardt F, Paeschke A, Rolfes M, et al. A database of german emotional speech[C]//9th *European Conference on Speech Communication and Technology*. Lisbon, Portugal, 2005:15－30.

[10] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. *Language Resources and Evaluation*, 2008, **42**(4): 335－359.

基于 Transformer 编码器的多级表示与融合特征输入的语音情感识别方法

贺正然[1]　沈起帆[1]　吴佳欣[2]　徐梦瑶[3]　赵　力[1]

([1] 东南大学信息科学与工程学院, 南京 210096)
([2] 东南大学微电子学院, 南京 210096)
([3]School of Computer Science and Software Engineering, University of Stirling, Stirling FK9 4LA, UK)

摘要:为了提高语音情感识别的准确度,探讨了将 Transformer 应用于语音情感识别的可能性.将对数梅尔尺度谱图及其一阶差分特征相融合作为输入,使用 Transformer 来提取分层语音表示,分析注意头个数和 Transformer 编码器层数的变化对识别精度的影响.结果表明,在 ABC、CASIA、DES、EMODB 和 IEMOCAP 语音情感数据库上,相比以 MFCC 为特征的 Transformer,所提模型的精度分别提高了 13.98%、8.14%、24.34%、8.16% 和 20.9%.该模型表现优于递归神经网络(RNN)、卷积神经网络(CNN)、Transformer 等其他模型.

关键词:语音情感识别;Transformer;多头注意力机制;融合特征

中图分类号:TP391.42