Speech emotion recognition based on crossmodal transformer and attention weight correction

Ryusei Terui* and Takeshi Yamada*
* University of Tsukuba, Japan
E-mail: s2320626@u.tsukuba.ac.jp

***Abstract*—In recent years, speech emotion recognition (SER) methods that use both acoustic features and text features derived through automatic speech recognition (ASR) have become mainstream. Furthermore, the crossmodal integration of acoustic and text features using a crossmodal transformer encoder has been proposed and succeeded in improving the SER accuracy. However, these methods have a problem in that ASR errors occur frequently, especially for speech that contains emotion, which affects the SER accuracy. To solve this problem, a method of correcting the self-attention weights based on the word-level confidence measure (CM), which indicates the reliability of ASR results, has been proposed. In this paper, we propose a method that combines the crossmodal transformer encoder and the attention weight correction with CM to further improve the SER accuracy. The network of the proposed method includes two different attention mechanisms: scaled dot-product attention and self-attention. In this paper, we applied the attention weight correction to each attention mechanism and verified their effectiveness. Results of experiments using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset revealed that the attention weight correction for self-attention can achieve higher SER improvement and the SER accuracy of the proposed methods is equal to or higher than that of state-of-the-art SER methods.**

I. Introduction

Emotion recognition is a key to realizing natural and smooth human-to-human and human-to-machine communication. Previously, many studies of the recognition of emotions from various modalities such as speech, facial expressions, gestures, and brain waves were conducted[1] [2] [3]. In particular, speech is a basic modality used on a daily basis and there are many situations where only speech can be used. Therefore, there are high expectations for speech emotion recognition (SER) technology. This technology makes it possible, for example, to provide responses that consider the client's emotions in call centers and responses that show empathy for users in speech assistant applications.

In previous SER methods, emotions have been classified using acoustic features based on frequency spectrograms such as the Mel-frequency cepstrum coefficient (MFCC), which is widely used in automatic speech recognition (ASR) and speaker recognition [1]. In recent years, methods that apply ASR to input speech to obtain the utterance text and classify emotions using both acoustic features and text features have become mainstream. Furthermore, the crossmodal integration of acoustic and text features using a crossmodal transformer encoder has been proposed and succeeded in improving the SER accuracy [4].

However, these methods have a problem in that ASR errors occur frequently, especially for speech that contains emotion, which affects the SER accuracy. To solve this problem, Feng *et al.* proposed a method of multi-task learning of ASR and SER to make ASR robust to emotional speech [5]. However, the computational and time costs are very high. On the other hand, Santoso *et al.* regarded ASR errors to be emotional cues and proposed a method of correcting the self-attention weights based on the word-level confidence measure [6], which indicates the reliability of ASR results, and showed its effectiveness [7].

In this paper, we propose a method that combines the crossmodal transformer encoder and the attention weight correction to further improve the SER accuracy. The network of our proposed method consists of a crossmodal transformer encoder for the crossmodal integration of acoustic and text features, a self-attention mechanism for emphasizing time frames important for classification, and a dense layer for classification. It includes two different attention mechanisms: scaled dot-product attention and self-attention. In this paper, we apply the attention weight correction to each attention mechanism and verify these effectiveness.

II. Conventional SER methods

A. Method using crossmodal transformer encoder

A method using a crossmodal transformer encoder was proposed to realize the crossmodal integration of acoustic features and text features. A transformer encoder is the encoder part of the transformer model [8]. It is composed of a multihead attention layer, a feed forward layer, and residual connections to these layers. Generally, the multihead attention layer of the singlemodal transformer encoder uses the same modality features for the query Q, the key K, and the value V. On the other hand, that of the crossmodal transformer encoder uses different modality features for K and V from Q. It has been reported that the SER accuracy is improved by performing the crossmodal integration of acoustic and text features instead of extracting them independently [4].

Fig. 1 shows an example of an SER method using a crossmodal transformer encoder. This is used as the base method in this paper. Fig. 2 also shows an overview of the crossmodal transformer encoder in the acoustic feature

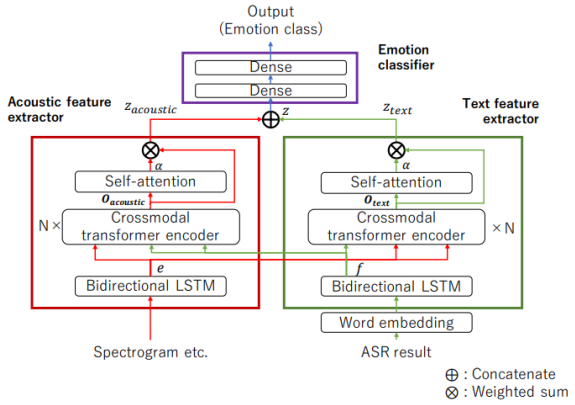

Fig. 1. Overview of the base method.

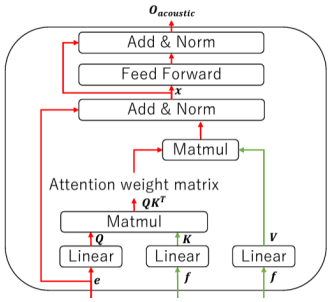

Fig. 2. Overview of the crossmodal transformer encoder.

extractor. In this figure, for simplicity, singlehead attention is used, but multihead attention is actually used. In the base method, the acoustic features are fed to a bidirectional long short-term memory (BLSTM) to extract the feature $\boldsymbol{e}_k$, which is defined for each time frame index k as

$$\boldsymbol{e}_k = \boldsymbol{g}_k \oplus \boldsymbol{h}_k, \tag{1}$$

where $\boldsymbol{g}_k$, $\boldsymbol{h}_k$, and $\oplus$ represent the forward and backward hidden states of the BLSTM and a concatenation operation, respectively. The text features are also fed to a BLSTM in the same manner as in the acoustic feature extraction to obtain $\boldsymbol{f}_l$ for the lth word. Then, $\boldsymbol{e}$ and $\boldsymbol{f}$ are fed to the crossmodal transformer encoder, and $\boldsymbol{e}$ is transformed to $\boldsymbol{Q}$ and $\boldsymbol{f}$ is transformed to $\boldsymbol{K}$ and $\boldsymbol{V}$ to apply the scaled dot-product attention as follows.

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

where d_k represents the dimension of $\boldsymbol{Q}$. This results in the crossmodal integration of acoustic and text features. We add $\boldsymbol{e}$ to the output of the scaled dot-product attention and normalize it to obtain $\boldsymbol{x}$, which is fed into the feed forward network layer defined as

$$FFN(\boldsymbol{x}) = max(0, \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2, \tag{3}$$

where $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{b}_1$, and $\boldsymbol{b}_2$ are trainable parameters. We add $\boldsymbol{x}$ to the output of the feed forward network layer and normalize

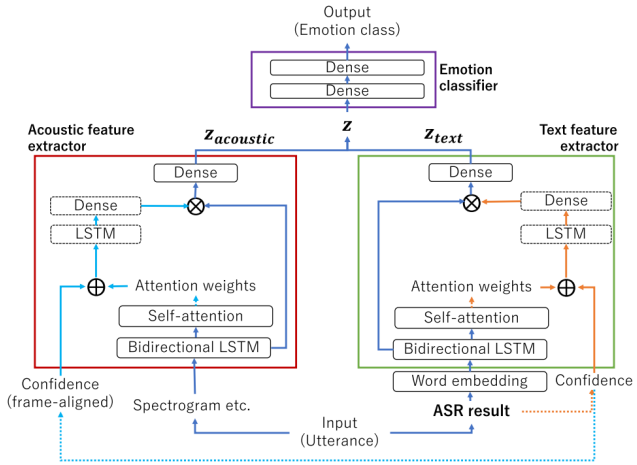

Fig. 3. Overview of the SER method with the attention weight correction.

it to obtain $\boldsymbol{O}_{acoustic}$, which is then fed into the self-attention mechanism to emphasize the important time frames.

$$\alpha_k = softmax(\boldsymbol{y}_k tanh(\boldsymbol{Y}\boldsymbol{O}^T_{acoustic,k})), \tag{4}$$

where α_k is the attention weight for the kth time frame, and $\boldsymbol{y}_k$ and $\boldsymbol{Y}$ are trainable parameters. The weighted sum $\boldsymbol{z}_{acoustic}$ from the attention weight α_k and $\boldsymbol{O}_{acoustic}$ is defined as

$$\boldsymbol{z}_{acoustic} = \sum_{k=1}^{T} \alpha_k \boldsymbol{O}_{acoustic,k}, \tag{5}$$

We apply the same processing as in the acoustic feature extraction to the text features to obtain $\boldsymbol{z}_{text}$. Finally, these two features are concatenated to obtain z and classified with a fully connected network to obtain the emotion class. Although it is superior in terms of the crossmodal integration of acoustic and text features, this method is vulnerable to ASR errors.

B. Method using attention weight correction

Fig. 3 shows an overview of the SER method with attention weight correction [7]. This method takes acoustic features such as a frequency spectrogram and the ASR text as input, and uses BLSTM to extract acoustic features and text features independently as in Eq. (1). Then, a feature vector that emphasizes important time frames (or words) is obtained using the self-attention mechanism. This is similar to Eqs. (4) and (5). The two features obtained in this manner are concatenated and fed into the fully connected layer to obtain the emotion class.

The advantage of this method is the attention weight correction. When ASR errors occur, unexpected attention weights may be added, which may lead to a decrease in SER accuracy. To solve this problem, a method of correcting the attention weights using a confidence measure (CM) output together with the utterance text from ASR was introduced. The CM is an indicator of the reliability of ASR results and is expressed as a value from 0 to 1, and the closer this value is to 1, the higher the reliability. Words with low reliability have a high possibility of ASR errors and are therefore likely to include

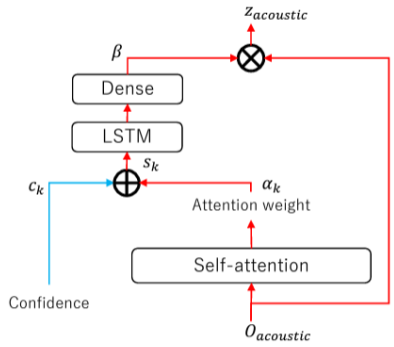

Fig. 4. Overview of the self-attention weight correction (proposed method 1).

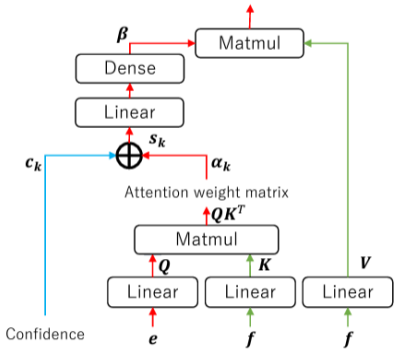

Fig. 5. Overview of the scaled dot-product attention weight correction (proposed method 2).

significant emotional features, so the attention weights of text features are corrected to be small and those of acoustic features are corrected to be large. The details of the attention weight correction method will be described later.

III. Proposed Method

In this section, we describe the proposed method in detail. The network of the proposed method is the same as that of the base method using the crossmodal transformer encoder in Fig. 1. The proposed method includes two different attention mechanisms, scaled dot-product attention and self-attention, but it is not clear how to apply attention weight correction to which attention mechanism. Therefore, in this paper, we compare the following two attention weight correction methods:

1) Self-attention weight correction (proposed method 1): Fig. 4 shows an overview of this method. This method is represented as

$$s_k = LSTM(\alpha_k \oplus c_k), \tag{6}$$

$$\beta_1, ..., \beta_T = softmax(s_1, ..., s_T), \tag{7}$$

where α_k is the attention weight for the kth time frame, c_k is the corresponding CM, and β_k is the corrected attention weight. This is the same method as the attention weight correction described in Sect. II.B.

2) Scaled dot-product attention weight correction (proposed method 2): Fig. 5 shows an overview of this method. It shows the network structure of the crossmodal transformer encoder in Fig. 1, and attention weight correction is applied to the attention weight matrix. First, the confidence matrix is obtained by calculating the cross product of CM vectors of acoustic features and text features. The corrected attention weight matrix is obtained as follows:

$$\mathbf{s_k} = Linear(\boldsymbol{\alpha_k} \oplus \mathbf{c_k}), \tag{8}$$

$$\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_T} = softmax(\mathbf{s_1}, ..., \mathbf{s_T}), \tag{9}$$

where $\boldsymbol{\alpha_k}$ is the kth column of the attention weight matrix, $\mathbf{c_k}$ is the column of the confidence matrix, and $\boldsymbol{\beta_k}$ is the column of the corrected attention weight matrix. This is a newly proposed method for scaled dot-product attention.

The proposed method 1 performs crossmodal integration while ignoring the presence of ASR errors, and then modifies the attention weights in the final self-attention step. On the other hand, the proposed method 2 performs crossmodal integration while considering the presence of ASR errors. In the next section, we compare the effectiveness of each method.

IV. Experiments

In this section, we first verify the effectiveness of crossmodal integration of acoustic features and text features. We then compare the effectiveness of the proposed method 1 with that of the proposed method 2, and compare their performance with that of state-of-the-art methods.

A. Experimental setting

In this study, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [9], which is widely used as one of the benchmarks in SER. This dataset consists of English speech data of 1 to 19 seconds from 10 speakers, 5 males and 5 females, and each utterance is labeled with an emotion class. In this paper, we use four emotion classes, namely, happy, sad, neutral, and angry, following past studies, whose numbers of data are 1689, 1084, 1708, and 1103, respectively.

The acoustic features are 33-dimensional vectors consisting of 20-dimensional MFCC, 12-dimensional constant Q-transform (CQT), and 1-dimensional fundamental frequency (F0). The text features are 768-dimensional word-embedding vectors obtained by pretrained BERT [10]. The utterance text is obtained using ASR pretrained with the LibriSpeech [11] dataset and the Kaldi speech recognition toolkit [12].

In this experiment, we use the following two evaluation measures: unweighted accuracy (UA), which is the classification accuracy for the entire test data, and weighted accuracy (WA),

TABLE I
SER PERFORMANCE OF THE BASE METHOD AND THE PROPOSED METHODS.

method	crossmodal integration	UA (%)	WA (%)
base method	w/o	73.4	75.7
base method	w	74.7	76.9
proposed method 1	w	75.4	77.5
proposed method 2	w	75.3	77.2

which considers the imbalance in the number of data among emotion classes in the test data. UA and WA are defined as

$$UA = \frac{\sum_{i=1}^{N} t_{ii}}{\sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij}}, \tag{10}$$

$$WA = \frac{1}{N} \sum_{i=1}^{N} \frac{t_{ii}}{\sum_{j=1}^{N} t_{ij}}, \tag{11}$$

where N is the number of emotion classes and t_{ij} is the number of data for class i that was classified as class j. In this experiment, we conducted fivefold cross validation and each evaluation value was the average of 5 validations. We used Adam [13] as the optimizer with a learning rate of 1.0e-4 and a weight decay of 1.0e-9. The loss function used for model training was the cross-entropy loss and the dropout and the batch size were set to 0.4 and 40, respectively. The results were taken from the best WA out of 100 epochs. Finally, the number of layers in the crossmodal transformer encoder was set to 3, and the number of heads in the multihead attention layer was set to 8.

B. Results

Table I shows the experimental results for each method. First, we confirm the effectiveness of crossmodal integration. In the table, the base method without crossmodal integration is the case where the same features are input to the Q, K, and V of the base method, and corresponds to extracting acoustic and text features independently. From the table, both the UA and WA of the base method are higher than those of the base method without crossmodal integration, confirming the effectiveness of crossmodal integration.

Then, we verify the effectiveness of the two types of attention weight correction. Compared with that of the base method, the WA values of the proposed methods 1 and 2 improved by 0.6% and 0.3%, respectively. This shows that the effect of ASR errors could be reduced by correcting the attention weights, considering the CM. In addition, the WA of the proposed method 1 is 0.3% higher than that of the proposed method 2. This implies that correcting the self-attention weight is more effective. One possible reason for this is that the process of self-attention weight correction is simpler and therefore easier to train.

Finally, Table II shows the UA and WA values of the proposed and state-of-the-art SER methods. These methods use acoustic and text features and parts of these methods adopt the crossmodal integration. We can see that the WA values of the proposed methods 1 and 2 are equal to or higher than those of the state-of-the-art SER methods.

TABLE II
SER PERFORMANCE OF THE STATE-OF-THE-ART AND PROPOSED METHODS.

method	crossmodal integration	UA (%)	WA (%)
Feng *et al.* [5]	w/o	69.7	68.1
Santoso *et al.* [7]	w/o	76.8	76.6
Chu *et al.* [4]	w	75.1	76.3
Zhang *et al.* [14]	w	76.4	77.1
Priyasad *et al.* [15]	w	76.8	77.3
Proposed method 1	w	75.4	77.5
Proposed method 2	w	75.3	77.2

From the above, it was confirmed that the combination of the crossmodal transformer encoder and the attention weight correction successfully improves the SER performance.

V. CONCLUSIONS

In this paper, we proposed a method that combines the crossmodal transformer encoder and the attention weight correction to further improve the SER accuracy. Since the network of the proposed method includes two different attention mechanisms, scaled dot-product attention and self-attention, we applied the attention weight correction to each of them and compared their effectiveness. The experimental results showed that the attention weight correction for self-attention can achieve higher SER accuracy and that the SER accuracy of the proposed methods is equal to or higher than that of state-of-the-art SER methods.

REFERENCES

[1] Z. Liao and S. Shen, "Speech emotion recognition based on swin-transformer," *Journal of Physics: Conference Series*, ser. 2508 012056, 2023.

[2] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[3] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[4] I.-H. Chu, Z. C. Xinlu, Y. M. Han, J. Xiao, and P. Chang, "Self-supervised cross-modal pretraining for speech emotion recognition and sentiment analysis," in *EMNLP*, 2022, pp. 5105–5114.

[5] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr," in *INTERSPEECH*, 2020, pp. 501–505.

[6] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[7] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, "Speech emotion recognition based on self-attention weight correction for acoustic and text features," *IEEE Access*, vol. 10, pp. 115 732–115 743, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture dataset," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, vol. 1, 2019, pp. 4171–4186.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio book," in *ICASSP*, 2015, pp. 5206–5210.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU*, 2011, pp. 1–4.

[13] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015, pp. 1–15.

[14] Z. Zhang, D. Liu, S. Liu, A. Wang, J. Gao, and Y. Li, "Turbo your multi-modal classification with contrastive learning," in *INTERSPEECH*, 2023, pp. 1848–1852.

[15] D. Prisayad, T. Fernando, S. Sridharan, S. Denman, and C. Fookes, "Dual memory fusion for multimodal speech emotion recognition," in *INTERSPEECH*, 2023, pp. 4543–4547.