# Speech emotion recognition based on the reconstruction of acoustic and text features in latent space

Jennifer Santoso*, Rintaro Sekiguchi*, Takeshi Yamada*,
Kenkichi Ishizuka†, Taiichi Hashimoto†, and Shoji Makino‡*
* University of Tsukuba, Japan
† RevComm Inc., Japan
‡ Waseda University, Japan
E-mail: j.santoso@mmlab.cs.tsukuba.ac.jp

***Abstract*—Speech emotion recognition (SER) has been actively studied in the recent decade and has achieved promising results. Most state-of-the-art SER methods are based on a classification approach that ultimately outputs the softmax probability over different emotion classes. On the other hand, we have recently introduced an anomalous sound detection approach to improve the SER performance of the neutral class. It uses a neutral speech detector consisting of an autoencoder that reconstructs acoustic and text features in latent space and is trained using only neutral speech data. The experimental result confirmed that the reconstruction error could be successfully used as a cue to decide whether or not the class is neutral and suggested that it could be applied to other emotion classes. In this paper, we propose an SER method based on the reconstruction of acoustic and text features in latent space, in which the reconstructor for different emotion classes, including the neutral class, is used. The proposed method selects the emotion class with the lowest normalized reconstruction error as the SER result. Unlike the classifier approach, one reconstructor is dedicated to each emotion class and trained using only the data of the target emotion class. Therefore, the reconstructor can be trained without being affected by imbalanced training data and also facilitates the application of data augmentation to only a specific emotion class. Our experimental result obtained using the IEMOCAP dataset showed that the proposed method improved the class-average weighted accuracy by 1.7% to 77.8% compared with the state-of-the-art SER methods.**

## I. Introduction

Emotion recognition, as a core component in human-to-human and human-to-machine interaction, has been an important field of study in recent years. The advancement of technology has enabled emotion recognition systems to receive various inputs, including speech, facial expressions, and biological signals. For example, one approach uses electroencephalograms to recognize emotions [1] whereas another uses sequences of facial expressions and physiological signals [2]. The combination of such information for emotion recognition has also been studied; for example, in one of the studies, speech, text, visual information, and motion capture are used and combined at the end for classification [3]. However, there are real-life situations, such as call center analysis and virtual voice assistants, where types of information other than speech are unusable. As one of the most common features used in emotion recognition, speech is rich in information on emotion and is readily available. This leads to the study of speech emotion recognition (SER), the task of recognizing a speaker's emotional state from their speech.

In the recent decade, state-of-the-art SER methods have achieved promising results through the use of various deep neural networks on speech information. One method uses convolutional neural networks to extract audio features, and the attention mechanism for importance weighing has been adopted as one of the benchmarks for SER [4]. The spread of automatic speech recognition systems (ASR) plays a major role in automatically obtaining text information from speech, enabling a combination of text and speech processing to improve SER methods. In several studies, SER combined with ASR results has further improved the SER performance. However, ASR performance degrades in the presence of emotions. To reduce the speech recognition error caused by emotions, the method that performs jointly SER training and ASR tuning has been proposed to be robust to emotional speeches [5] and the use of the confidence measure (CM) outputted together with the ASR text has been proposed to mitigate the speech recognition errors, resulting in a better SER performance [6]. Using more advanced deep-learning architecture such as the transformers, some groups have further raised the efficiency of feature extraction of emotional speeches by focusing on emotion-related information [7] and using the correlation between speech and text information [8] [9]. Most methods are based on a classification approach, which outputs the softmax probability of different emotion classes.

One limitation of the classification approach is the need to balance the training data since otherwise, it would result in a classifier biased toward a certain class. The performance of the emotion class with low performance can be improved by increasing the training data for that class. However, it would be difficult to maintain the balance of the training data. Another limitation is that in the case of additional emotional classes, it would be more difficult to add new emotion classes or to retrain the classifier from scratch.

978-616-590-477-3 ©2022 APSIPA

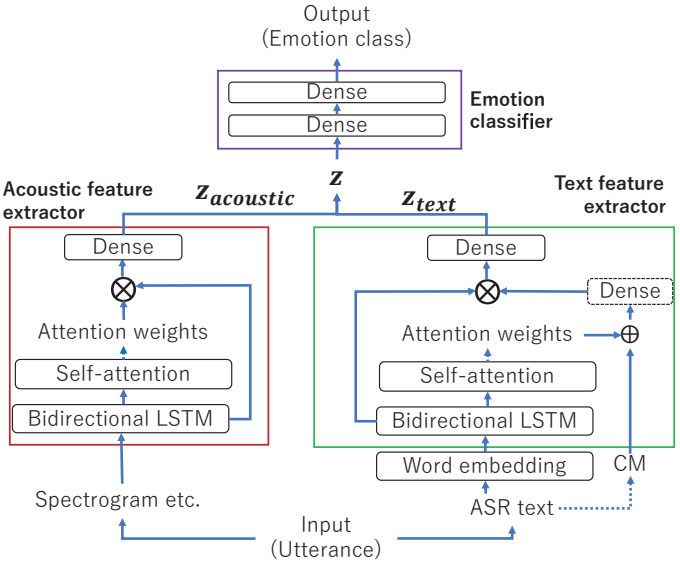


Fig. 1. Architecture of base SER method

On the other hand, we introduced an anomaly detection-based approach to the SER domain in our previous study [10], where we used a deep autoencoder to reconstruct the acoustic and text features in the latent spaces from a pretrained SER method; these features are compact representations of emotional speeches. The deep autoencoder is trained only on neutral speeches and is used to decide whether a speech is neutral or emotional on the basis of a decision threshold. The experimental result showed performance improvement in neutral speech and confirmed that the reconstruction error could be used as an indicator to decide whether a speech is neutral or emotional.

In this study, we propose an SER method based on reconstruction error. First, we extract the acoustic and text features in latent space by using a pretrained SER classifier [6]. Second, the extracted features are fed into the reconstructor for each emotion class. Finally, the emotion class is judged to be that with the lowest normalized reconstruction error. The main advantage of our proposed method is the possibility of training the autoencoder separately for each emotion class, therefore alleviating the need for data balancing. Furthermore, the data augmentation of the latent space can be done specifically for each emotion class without being affected by the others. We conduct experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [11] to verify the effectiveness of this approach. Finally, we compare the performance characteristics of our proposed method with those of the state-of-the-art SER classification methods.

## II. Overview of Base SER Method

Fig. 1 illustrates the SER method, which is similar to that in our previous work [6]. The SER method consists of three main parts: the acoustic feature extractor, the text feature extractor, and the emotion classifier.

The input acoustic feature in the base SER method consists of Mel-frequency cepstrum coefficients (MFCCs), constant Q-transform (CQT), and fundamental frequency (F0). These features are then fed to the bidirectional LSTM (BLSTM) [12] network to obtain $\mathbf{e}_i$, which is defined as

$$\mathbf{e}_i = \mathbf{g}_i \oplus \mathbf{h}_i, \tag{1}$$

where $\mathbf{g}_i$, $\mathbf{h}_i$, and $\oplus$ represent the forward hidden states of BLSTM, the backward hidden states of BLSTM, and concatenation, respectively. $\mathbf{e}_i$ is then weighed for its importance by the self-attention mechanism [13] defined as

$$\alpha_i = softmax(\mathbf{w}_i tanh(\mathbf{W}\mathbf{e}_i^T)). \tag{2}$$

Here, $\alpha_i$ is the attention weight for each frame, and $\mathbf{w}$ and $\mathbf{W}$ are trainable parameters that represent a fully connected layer. Therefore, the weighted sum $\mathbf{v}$ from BLSTM and attention weights is defined as

$$\mathbf{v} = \sum_{i=1}^{T} \alpha_i \mathbf{e}_i. \tag{3}$$

After the weighted sum $\mathbf{v}$ is calculated, it is fed to a single fully connected layer to obtain a fixed-length latent space, $\mathbf{z}_{acoustic}$, of acoustic features.

The text feature extraction in the base SER method uses the ASR text from the same speech data as the input. The ASR result is first encoded by BERT word embedding [14], in which the resulting features are fed to the text feature extractor. The text feature extractor is based on BLSTM and a self-attention mechanism similar to that used in the acoustic feature extractor. Furthermore, an additional self-attention weight correction that uses the confidence measure (CM), a metric that indicates the reliability of ASR results, was introduced in our previous study [6]. As emotions adversely affect the ASR performance and cause speech recognition errors, CM acts as another weight that helps mitigate the effects of speech recognition errors from the ASR text. Combining the self-attention mechanism weight and CM yields more precise weights for textual information. In the self-attention weight correction, the CM $c_i$ is concatenated with the attention weight $\alpha_i$, which is then fed to a fully connected layer and normalized using the softmax function to obtain the updated attention weights. The self-attention weight correction is expressed as

$$\beta_i = softmax(\mathbf{W}'(\alpha_i \oplus c_i)), \tag{4}$$

$$\mathbf{z}_{text} = \sum_{i=1}^{N} \beta_i \mathbf{e}_i. \tag{5}$$

The fully connected layer represented by the trainable parameter $\mathbf{W}'$ learns and adjusts the attention weights by considering the CM aligned to the same word position. The resulting self-attention weight $\beta_i$ is then used to calculate the weighted sum of BLSTM outputs, producing a new fixed-length latent representation from the text $\mathbf{z}_{text}$.

The emotion classifier receives the latent representation $\mathbf{z} = \mathbf{z}_{acoustic} \oplus \mathbf{z}_{text}$, which is then fed to layers of a fully

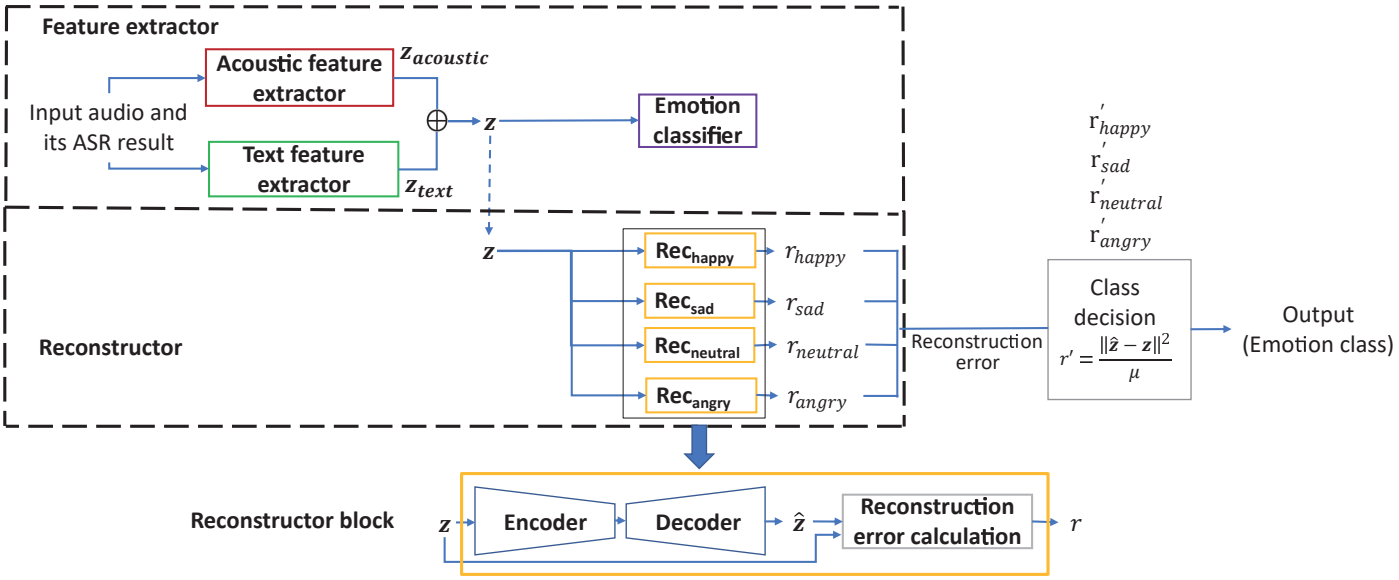


Fig. 2. Proposed method flow

connected network, resulting in the softmax probability over different emotion classes. The emotion class is selected with the highest probability among the different emotion classes.

## III. Proposed Method

### A. Overview

Fig. 2 illustrates our proposed method flow. Our proposed method consists of the feature extractor, the reconstructor for each emotion class, and a class decision. The feature extractor is taken from the acoustic and text feature extractor of the pretrained SER method explained in Sect. II, resulting in the latent representation $\mathbf{z}$. This is then fed to the autoencoder-based reconstructor for each target emotion class. Each reconstructor reconstructs $\mathbf{z}$, resulting in the reconstructed feature vector $\hat{\mathbf{z}}$ and having the calculated reconstruction error. We select the emotion class with the lowest normalized reconstruction error.

### B. Reconstructor

Our proposed reconstructor has the architecture of a deep autoencoder [15] and is made of an encoder and a decoder, both in the form of two neural networks. The autoencoder is mainly effective in dimensionality reduction and the representation of higher-dimensional data. The use of an autoencoder in our proposed method was inspired by the anomaly detection-based approach applied in the anomalous sound detection task [16], which was previously handled using a classification-based approach. In anomalous sound detection, the autoencoder is used to reconstruct the spectrogram and detect the anomalous machine sound. The success of anomalous sound detection has made the autoencoder a commonly used solution for reconstruction and detection tasks. However, applying this approach to the SER domain has several challenges. First, reconstructing a spectrogram for speech is difficult owing to the high dimensions and variable lengths. Moreover, it is necessary to keep the textual information, providing cues to the target emotion.

One possible idea to solve these problems is to use latent representations as the reconstruction target for the autoencoder. The latent representation is usually taken from a pretrained method, has a fixed length, and contains a compact representation of the input features important to the task of the pretrained method. Using the autoencoder on the latent space has been proven effective in improving the anomaly detection performance in image anomaly detection [17]. Following the success of image anomaly detection, we introduced the use of the autoencoder and latent space to both our previous method and our proposed method.

In our reconstructor, the autoencoder learns the representation of speeches in a target emotion class from the latent representation $\mathbf{z}$ of the pretrained SER method. The main strength of our reconstructor is the ability to learn a more specific representation of the acoustic and textual information $\mathbf{z}$ as the reconstruction target. $\mathbf{z}$ is a compact representation of features prominent in SER, enabling easy reconstruction. $\mathbf{z}$ is transformed into a bottleneck representation $\mathbf{v}$ with the encoder $\mathcal{E}$, whereas the decoder $\mathcal{D}$ maps back the bottleneck representation into the reconstructed latent representation $\hat{\mathbf{z}}$. The process is defined as

$$\mathbf{v} = \mathcal{E}(\mathbf{z}|\theta_E), \tag{6}$$

$$\hat{\mathbf{z}} = \mathcal{D}(\mathbf{v}|\theta_D), \tag{7}$$

where $\theta_E$ and $\theta_D$ represent the parameter set of an encoder and a decoder, respectively. The reconstruction error of the reconstructor is computed using the mean square error (MSE)

$$r = \|\mathbf{z} - \hat{\mathbf{z}}\|^2, \tag{8}$$

where $dim$ is the dimension of $\mathbf{z}$.

### C. Class decision

In anomaly detection tasks, it is common to determine whether data is anomalous or not on the basis of the decision threshold. For example, the threshold used for the autoencoder in our previous study [10] is obtained from the Gamma distribution [18] percentile value of the training data. If the anomaly score exceeds the decision threshold, the data is considered anomalous and vice versa. It is possible to use the decision threshold to decide whether the reconstructor result corresponds to the target emotion class or not. However, integrating reconstructor results from each emotion class obtained on the basis the decision threshold would raise several issues, such as the difficulty in determining the optimum decision threshold value for each reconstructor and classifying a speech into exactly one emotion class.

To solve these problems, we propose integrating the reconstructor results from each emotion class and determining the emotion class of a speech without using a decision threshold. First, we calculate $r'$, which is the normalized reconstruction error, defined as

$$r' = \frac{\|\mathbf{z} - \hat{\mathbf{z}}\|^2}{\mu}, \tag{9}$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{z}_n - \hat{\mathbf{z}_n}\|^2, \tag{10}$$

where $\mu$ represents the average reconstruction error of the training data for the target emotion class, $z$ and $\hat{z}$ represent the acoustic and text features in the latent space and its reconstruction version, and $N$ represents the number of data in the training set. Then, we calculate $r'$ for each target emotion class. Finally, we select the emotion class with the smallest $r'$ as the final result.

## IV. Experiments

### A. Overview

The aim of the experiment is to examine the effectiveness of the proposed method in improving SER performance. We compare the unweighted and weighted accuracy and the F-score of each emotion class with those obtained from the state-of-the-art SER methods.

### B. Dataset

To evaluate our proposed method, we use the IEMOCAP dataset [11], one of the benchmark datasets for emotion recognition. The IEMOCAP dataset consists of scripted and improvised emotional speeches divided into five sessions, each containing one male and one female speaker. There are ten speakers (five males and five females) in the IEMOCAP dataset.

The pretrained SER method was trained using the data from four classes (happy, sad, neutral, and angry). We included the utterances labeled as excited with the utterances labeled as happy to make the dataset condition similar to that in previous works. The experiments were performed in five-fold cross-validation. The training set comprises four sessions, and the test set comprises the remaining one session to ensure speaker independence. The F-scores reported are based on the combined results from all five folds, not from averaging the F-score in each fold. The details of the dataset are shown in Table 1.

TABLE I
DATASET SPECIFICATIONS

| | | |
|---|---|---|
| Dataset | IEMOCAP | |
| Speakers | 5 males and 5 females | |
| Utterance length | 1−19 s | |
| Number of utterances | Happy | 1689 |
| | Sad | 1084 |
| | Neutral | 1708 |
| | Angry | 1103 |

The reconstructor in the proposed method was trained using the same five fold cross-validation setting as the pretrained SER method. For each emotion class, we train the reconstructor separately to learn the data representation of the respective emotion class. Therefore, the training set for the reconstructor for each emotion class contains only the speeches labeled as the trained emotion class from each of the four sessions. On the other hand, the test set of each reconstructor uses the same dataset as that for the SER.

### C. Input features

The features inputted to the pretrained SER method were divided into two parts: acoustic feature extraction and textual feature extraction [6]. For acoustic feature extraction, we extracted a 33-dimensional feature consisting of 20-dimensional MFCCs, 12-dimensional CQT, and one-dimensional F0. All of the acoustic features are extracted using Librosa [19]. For the textual features, first, we conducted ASR on the input speeches using a recognizer pretrained with the Librispeech [20] dataset and Kaldi speech recognition toolkit [21]. Librispeech consists of approximately 1000 hours of speech sampled at 16 kHz. As a reference, the word error rate of the ASR tested on the Librispeech dataset is 3.8%, whereas the word error rate of the ASR tested on the IEMOCAP dataset is 43.5%. Next, we encoded the ASR texts using BERT [14] pretrained using lower-case English texts. The pretrained BERT consists of 12 layers and 110M parameters, resulting in 768-dimensional textual features.

### D. SER method and reconstructor specifications

The pretrained SER method consists of an acoustic feature extractor, a textual feature extractor, and an emotion classifier. The specifications for the pretrained SER method are the same as the one used in the previous study [6]. The feature extractor for both the acoustic and text features used one layer of BLSTM with 128 units, a self-attention mechanism with 128 units for the acoustic feature extractor, and the additional CM-based correction mechanism for the text feature extractor. The resulting latent representation $\mathbf{z}$ from the base SER method is a 256-dimensional vector consisting of a 128-dimensional vector from each of the acoustic and text features. In the pretrained SER method, $\mathbf{z}$ is fed to an emotion classifier consisting of a

TABLE II
SER PERFORMANCE CHARACTERISTICS (UA, WA) OF OUR PROPOSED METHOD AND STATE-OF-THE-ART METHODS. THE SYMBOL '–' MEANS THAT THE VALUE IS NOT DESCRIBED IN THE PAPER.

| Method | UA (%) | WA (%) |
|---|---|---|
| Neumann and Vu [4] | – | 56.1 |
| Feng et al. [5] | 69.7 | 68.6 |
| Chen et al. [7] | 75.3 | 74.3 |
| Siriwardhana et al. [8] | 75.5 | – |
| Base SER method [6] | 75.9 | 76.1 |
| Wang et al. [9] | 77.1 | 76.8 |
| Proposed method | **77.8** | **77.8** |

TABLE III
SER PERFORMANCE CHARACTERISTICS (F-SCORE) OF OUR PROPOSED METHOD AND STATE-OF-THE-ART METHODS. THE SYMBOL '–' MEANS THAT THE VALUE IS NOT DESCRIBED IN THE PAPER.

| Method | F-score (%) | | | |
|---|---|---|---|---|
| | Happy | Sad | Neutral | Angry |
| Neumann and Vu [4] | 58.2 | 51.9 | 52.8 | 66.5 |
| Feng et al. [5] | 69.1 | 70.5 | 61.0 | 77.3 |
| Chen et al. [7] | – | – | – | – |
| Siriwardhana et al. [8] | 77.1 | 78.4 | 64.7 | **81.9** |
| Base SER method [6] | 81.5 | 76.2 | 67.4 | 80.4 |
| Wang et al. [9] | – | – | – | – |
| Proposed method | **84.7** | **79.3** | **71.6** | 74.3 |

fully connected network with (256–64–4) units. The output of the pretrained SER method is assigned to the softmax probability of the four emotion classes, where the highest probability identifies the final emotion class. The pretrained SER method uses softmax cross-entropy as the loss function. The optimizer is set to Adam [22] with a learning rate of 0.0001 and a dropout of 0.2.

The reconstructor for each emotion class is an autoencoder consisting of nine layers with (256–128–64–32–16–32–64–128–256) units. The autoencoder is trained with mean squared error as the loss function. The optimizer is set to Adam with a learning rate of 0.00001 and a dropout of 0.2. We evaluate the SER performance using the average unweighted accuracy (UA), average weighted accuracy (WA), and F-score of each emotion class. The pretrained SER method used as the feature extractor was taken from the highest WA of the test data out of 100 epochs. In addition, the reconstructor for each emotion class, which is later integrated into our proposed method, was taken from the highest F-score for the emotion class of the test data out of 200 epochs.

### *E. Results*

Table 2 shows the UA and WA of our proposed method and state-of-the-art SER methods. Our proposed method achieved UA and WA of 77.8% and 77.8%, respectively. These results indicate UA and WA improvement of 1.9% and 1.7% over the base SER method, which achieved UA and WA of 75.9% and 76.1%, respectively. Our proposed method outperformed the state-of-the-art SER methods in terms of UA and WA.

Table 3 shows the F-scores of our proposed method and state-of-the-art SER methods. Overall, the F-score shows the effectiveness of our proposed method of SER based on reconstruction error, which is superior to the F-score of happy, sad, and neutral classes of the base SER method by 3.1%–4.2%. Fig. 3 shows the confusion matrices for the base SER method and our proposed method. Similar to the result of the F-score, our proposed method shows an increased performance for the happy class, sad class, and neutral classes in terms of accuracy. On the other hand, there is a deterioration in the performance for the angry class, with many of the speeches being incorrectly recognized as a neutral class. These results imply that the classifier-based SER methods and our proposed method have different strengths regarding performance for different emotion classes. Therefore, integrating our proposed method with the classifier-based SER methods would potentially boost the SER performance.

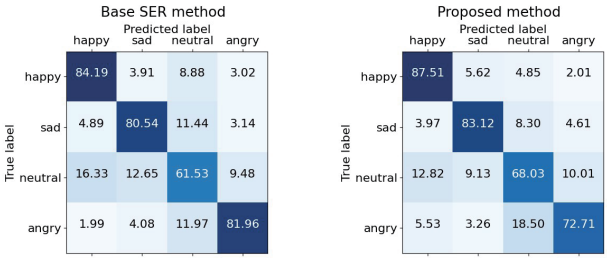


Fig. 3. Confusion matrices (in %) for base SER method and our proposed method

## V. CONCLUSIONS

In this paper, we proposed an SER method based on the reconstruction of acoustic and text features in latent space. The proposed method selects the emotion class with the lowest normalized reconstruction error as the SER result. The main advantage of our method is the possibility of training the reconstructor separately for each emotion class. The experimental results confirmed that our proposed method based on the reconstruction approach improves the overall SER performance by 1.9% on the UA and 1.7% on the WA for the IEMOCAP dataset, slightly outperforming most state-of-the-art methods based on the classification approach. In addition, the proposed method improved the SER performance for most emotion classes in terms of F-score. In future work, we will investigate how to integrate our proposed method with classifier-based SER methods to further improve the SER performance.

## ACKNOWLEDGMENTS

Part of this work was supported by JST SPRING, Grant Number JPMJSP2124.

## REFERENCES

[1] X. Li, W. Zheng, Y. Zong, H. Chang, C. Lu, "Attention-based spatio-temporal graphic LSTM for EEG emotion recognition," Proc. IJCNN, pp. 1–8, 2021.
[2] Y. Ouzar, F. Bousefsaf, D. Djeldjli, C. Maaoui, "Video-Based multimodal spontaneous emotion recognition using facial expressions and physiological signals," Proc. CVPR, pp. 2460–2469, 2022.
[3] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intelligent Systems, vol. 33, no. 6, pp. 17-25, 2018.

[4] M. Neumann, N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," Proc. Interspeech, pp. 1263–1267, 2017.
[5] H. Feng, S. Ueno, T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word ASR," Proc. Interspeech, pp. 501–505, 2020.
[6] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, T. Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure," Proc. Interspeech, pp. 1947–1951, 2021.
[7] W. Chen, X. Xing, X. Xu, J. Yang, J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," Proc. ICASSP, pp. 6897–6901, 2022.
[8] S. Siriwardhana, A. Reis, R. Weerasekera, S. Nanayakkara, "Jointly fine-tuning "BERT-like" self-supervised models to improve multimodal speech emotion recognition," Proc. Interspeech, pp. 3755–3759, 2020.
[9] Y. Wang, G. Shen, Y. Xu, J. Li, Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition," Proc. Interspeech, pp. 4518–4522, 2021.
[10] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, S. Makino, "Performance improvement of speech emotion recognition by neutral speech detection using autoencoder and intermediate representation," Proc. Interspeech, 2022. (in press)
[11] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture dataset," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.
[12] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," Proc. ICANN, vol. 2, pp. 799–804, 2005.
[13] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, "A structured self-attentive sentence embedding," Proc. ICLR, 2017.
[14] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL, pp. 4171–4186, 2019.
[15] D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning internal representations by error propagation," Parallel Distributed Processing. Vol 1: Foundations, chapter 8, pp. 318-–362. MIT Press, Cambridge, 1986.
[16] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma," IEEE/ACM Transactions on Audio Speech and Language Processing, pp. 212–224, 2019.
[17] V. Abdelzad, K. Czarnecki, R. Salay, T. Denounden, S. Vernekar, B. Phan. "Detecting out-of-distribution inputs in deep neural networks using an early layer output." arXiv preprint arXiv:1910.10307, 2019.
[18] K. O. Bowman, L. R. Shenton, "Gamma Distribution," International Encyclopedia of Statistical Science, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 573–575, 2011.
[19] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, "librosa: Audio and music signal analysis in Python," Proc. Python in Science Conference, vol. 8, pp. 18–25, 2015.
[20] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "Librispeech: an ASR corpus based on public domain speech books," Proc. ICASSP, pp. 5206–5210, 2015.
[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovský, "The Kaldi speech recognition toolkit," Proc. ASRU, pp. 1–4, 2011.
[22] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.