IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 11, 2021, accepted February 12, 2021, date of publication March 19, 2021, date of current version May 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3067460

Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset

MINGKE XU[1], FAN ZHANG[2], AND WEI ZHANG[3]
[1]Nanjing Tech University, Nanjing 211816, China
[2]IBM Data and AI, Boston, MA 01460, USA
[3]IBM Research AI, Boston, MA 01460, USA

Corresponding author: Fan Zhang (fzhang@us.ibm.com)

ABSTRACT Speech Emotion Recognition (SER) refers to the use of machines to recognize the emotions of a speaker from his (or her) speech. SER benefits Human-Computer Interaction(HCI). But there are still many problems in SER research, *e.g.*, the lack of high-quality data, insufficient model accuracy, little research under noisy environments, *etc.* In this paper, we proposed a method called Head Fusion based on the multi-head attention mechanism to improve the accuracy of SER. We implemented an attention-based convolutional neural network(ACNN) model and conducted experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) data set. The accuracy is improved to 76.18% (weighted accuracy, WA) and 76.36% (unweighted accuracy, UA). To the best of our knowledge, compared with the state-of-the-art result on this dataset (76.4% of WA and 70.1% of WA), we achieved a UA improvement of about 6% absolute while achieving a similar WA. Furthermore, We conducted empirical experiments by injecting speech data with 50 types of common noises. We inject the noises by altering the noise intensity, time-shifting the noises, and mixing different noise types, to identify their varied impacts on the SER accuracy and verify the robustness of our model. This work will also help researchers and engineers properly add their training data by using speech data with the appropriate types of noises to alleviate the problem of insufficient high-quality data.

INDEX TERMS Speech emotion recognition, convolutional neural network, attention mechanism, noise reduction.

I. INTRODUCTION

Emotion recognition plays an important role in Human–Computer Interaction(HCI). With the development of deep learning technology, it has become possible to recognize human emotions in terms of speech [1]–[7], text [8], [9], and facial [10], [11]. Speech is an important carrier of human communication, and it makes sense to recognize emotions from speech. Speech Emotion Recognition(SER) has wide application prospects on psychological assessment [12], robots [13], mobile services [14], *etc*. For example, a psychologist formulates a treatment plan according to the emotions hidden in the patient's speech.

Emotions can be defined by discrete models and dimensional models. Although it is intuitive to define emotions with discrete labels like 'happy', 'angry', 'sad', *etc.*, there are about 300 emotional states [15] in real life and it is hard to classify so many emotions. Therefore, discrete models define emotions as several basic emotions [16], and express more emotions through a combination of the basic emotions. So we only need to classify basic emotions [17]. On the other hand, dimensional models describe emotional states as points in a multidimensional emotional space [18]. Most of the current SER researches are based on discrete emotion models, *i.e.*, mapping utterances to discrete emotion tags.

Deep learning has accelerated the progress of recognizing human emotions from speech, but there are still deficiencies in the research of SER, such as data shortage and insufficient model accuracy.

Unlike Automatic Speech Recognition(ASR), current SER research faces the problem of lacking high-quality data. Generally speaking, it is expensive to invite actors to perform emotions in the studio and it is easier to collect noisy data in real life. On the other hand, emotions are subjective and

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Shen.

influenced by many factors, language and culture have an important influence on the judgment of emotions in speech [19], which increases the cost of data labeling. In addition, improvised speech is easier to be recognized than those performed in the studio according to the script, which has been demonstrated in previous research [20], [21].

Also due to the subjectivity and uncertainty of emotions, for machines, it is still a huge challenge to recognize emotional content in human speech, *i.e.*, the current SER model still does not accurately recognize human emotions.

In addition, due to the influence of the environment, speech sounds are often mixed with a variety of noises, so research on speech emotion in the noisy environment is practically significant. In real-life applications, it is difficult to acquire speech data to be recognized without noise interference.

Therefore, we are committed to improving the accuracy of the SER model and exploring the impact of noise. In this paper, our main contributions are as follows:

- We proposed a method called Head Fusion based on multi-head self-attention and designed an attention-based convolutional neural network(ACNN) model. The model improves accuracy to 76.18% (Weighted Accuracy, WA) and 76.36% (Unweighted Accuracy, UA) on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) data set, which is state-of-the-art.
- We study the impact of the noise intensity, time-shifting effect, and found out that the intensity influences the accuracy significantly but it does not change the accuracy relationship between using different noise, and the time-shifting slightly influences the accuracy.
- We found that in case of lacking training data, using speech data mixed with a select-few category of noise can effectively improve the accuracy of SER. We classified 50 different types of noises according to their characteristics and presented the empirical discovery of the appropriate noise types suitable for additional training data.

The organization of this paper is as follows. Section II describes some existing related work in the field of SER. Section III describes the data set researched in our paper. Section IV describes the method and model architecture proposed in our paper. Section V describes the experimental setup, results, and related analysis. Section VI describes the conclusion of our paper and the future work for SER.

II. RELATED WORK

Before the era of deep learning, for SER, researchers mostly use complex hand-crafted features (*e.g.* IS09, eGeMaps, *etc.*) and traditional machine learning methods (*e.g.* HMM, SVM, *etc.*) [22], [23]. Deep learning accelerated the progress of SER research. Han *et al.* [1] first applied deep learning to SER in 2014. Chen *et al.* [2] combines convolutional neural networks (CNN) and Long Short-Term Memory(LSTM) in their SER model. Wu *et al.* [3] replace traditional CNN with capsule networks (CapsNet) to improve the accuracy of SER. Xu *et al.* [4] use Gate Recurrent Unit(GRU) to calculate features from frame level and utterance level respectively for recognition. However, using ordinary deep learning algorithms without attention cannot achieve good results.

Since Google has greatly improved the accuracy of machine translation [24], the attention mechanism is being used more and more in deep learning. In speech recognition, attention mechanism has been used in many tasks such as ASR [25], speaker recognition [26] and SER [2], [20], [21], [27], also in our work.

In addition, some studies combine speech and text for recognition. For example, S. Yoon *et al.* use ASR to obtain text, use an attention mechanism to calculate speech-related parts from the text, and use BLSTM for recognition [28]. However, a text combined method depends on the accuracy of ASR.

Relatively speaking, there are still few studies introducing noise environment. Researchers usually use front-end noise reduction [29], feature set design [30], multimodal recognition [31] and *etc.* to deal with noise. Their research applied fewer types of noise or even only the white noise. Their primary focus is mostly on the noise robustness or noise reduction, instead of studying the impact of various noises themselves.

The above research is based on the discrete emotion model. There is also some research based on the dimensional emotion model [32], [33]. However, the definition of the dimensional model is still unclear and which dimensions to use for research are still controversial. Therefore, it has not yet become the mainstream research method for SER.

In our work, we use audio without text for recognition based on the discrete emotion model. We use CNN combined with self-attention as the main structure of our model, and the attention part is modified for higher accuracy. We carried out SER research on 50 types of noise, and we altered noise injection with many different approaches. We conclude with solutions purely experiment-based, aiming at properly assisting people augment the data for SER research.

III. DATA

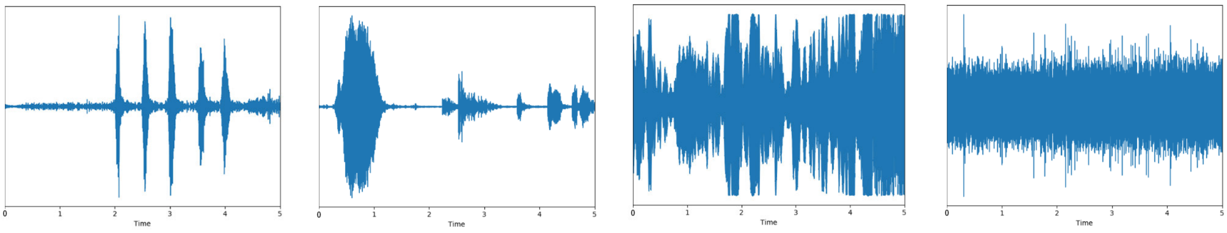
A. IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) data set [34] was proposed in 2008 and widely used as an SER benchmark since then. The total duration is about 12 hours of 5 parts and each part is a dialogue between two actors. Each part is a scene where an actor and an actress talk to each other. In total there were ten actors recruited from the Drama Department of University of Southern California.

IEMOCAP contains nine classes of emotions (anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state). For each utterance, there are more than three evaluators. Only when more than half of the evaluators have the same opinion, the utterance will be marked with the corresponding label, otherwise, it will be labeled as ‘other’. Obtaining data in this way is relatively natural.

TABLE 1. The four categories of the noises in the ESC-50.

Category	Separation	Fluctuations	Examples
Discrete-Short(D-S)	Obvious	Obvious	Door knocking, footsteps, dog barking, glass breaking, coughing *etc.*
Discrete-Long(D-L)	Obvious	Not Obvious	Cow mooing, car horns, baby crying, snoring, door - wood creaks, *etc.*
Continuous-Short(C-S)	Not Obvious	Obvious	Clock alarms, sirens, clapping, brushing, church bell ringings, *etc.*
Continuous-Long(C-L)	Not Obvious	Not Obvious	Rain, wind, helicopter, engine, chainsaw, *etc.*



(a) D-S (b) D-L (c) C-S (d) C-L

FIGURE 1. Examples of waveforms for four categories of noise. (a) (d) are examples of noise waveforms of D-S, D-L, C-S, and C-L respectively. In (a), the sound can be clearly divided into several segments, and each segment of the sound quickly drops after reaching a high point. In (b), the sound can also be clearly divided into several segments, but each segment of the sound slowly drops after reaching a high point. In (c), the sound is a continuous segment and there are obvious fluctuations. In (d), the sound is also a continuous segment, but there is no obvious fluctuation.

However, the utterances are labeled by evaluators rather than purposely designed, which causes is a large imbalance of emotion occurrences in the IEMOCAP data set. Since excitement and happiness have a certain degree of similarity and there are too few happy utterances, researchers sometimes replace happiness with excitement or combine excitement and happiness to increase the amount of data [21], [27], [35].

Besides, the IEMOCAP data set is also divided into scripted and improvised parts according to whether the actors read scripts. The recognition accuracy of the improvised part is higher in the SER task in general because actors can focus more on the emotions they need to perform rather than reciting scripts [20], [21].

Since IEMOCAP is of high quality and has a large amount of data, and there are many studies on it, we use IEMOCAP as the main data set for our research. In this paper, following other published work, we choose four representative emotions in the improvised part, *i.e.*, anger (1103 utterances), sadness (1084 utterances), excitement (1041 utterances), and neutral state (1708 utterances).

B. RAVDESS

The RAVDESS data set [36] contains two emotional parts-speech part and song part. We mainly focus on its speech part, which has 1440 utterances acted by 24 professional actors (12 female, 12 male). There are 8 classes of emotion contained in the speech part (calmness, happiness, sadness, anger, fear, surprise, disgust, and neutral state). Excluding the neutral state, other classes of emotion include two intensities (normal and strong). In the process of collecting data, actors are asked to speak two fixed sentences with different classes of emotion. Therefore, the RAVDESS data set has a good balance, but relatively poor naturalness.

We use the RAVDESS data set as a secondary data set and take all 8 classes of emotion into research for its balance.

C. ESC-50

The Environmental Sound Classification (ESC-50) data set [37] contains 50 types of environmental noise and each type of noise contains 40 audio snippets. Each audio snippet lasts for 5 seconds. We use the data set as the source of the noise.

In the ESC dataset, noise is divided into 5 categories according to its source. But in our work, We reclassify them into 4 categories based on whether there are obvious separations in between noise segments and whether there are obvious and intermittent fluctuations. As shown in Table 1, we name these 4 categories Discrete-Short(D-S), Discrete-Long(D-L), Continuous-Short(C-S) and Continuous-Long(C-L) respectively. FIGURE 1 shows examples of waveforms for the four categories of noise. In this paper, we use 'type' to represent a single type of noise(*e.g.* Door knocking, Cow mooing, *etc*) and use 'category' to represent a set of types of noise(*i.e.* D-S, D-L, C-S, and C-L).

IV. METHODOLOGY

A. FEATURE

We use Mel-scale Frequency Cepstral Coefficients (MFCCs) as input, an audio feature that is widely used in the field of speech recognition.

First, we use a Hanning window with a length of 2048 and a hop length of 512 to perform a short-term Fourier transform (STFT) on the audio signal and obtain the power spectrogram of the audio signal. Then we use mel filters to map the spectrogram to Mel-scale and take the logs to obtain the log Mel- spectrogram. Finally, we use a discrete cosine transform (DCT) transformation to obtain MFCCs.

We use the Librosa audio processing library [38] to implement the MFCCs extraction.

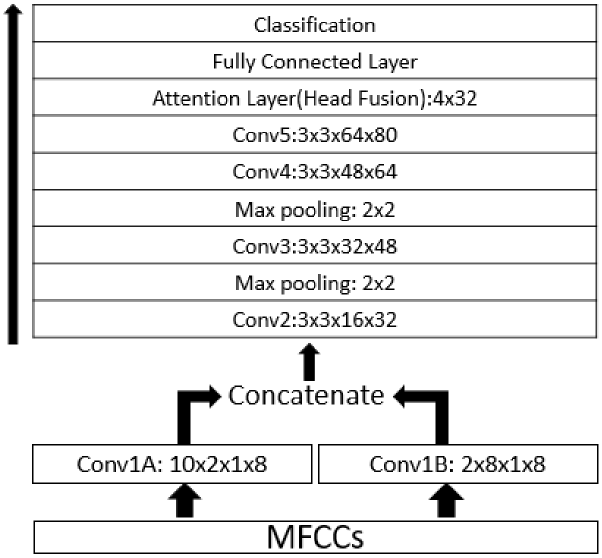


FIGURE 2. **Model Architecture The model contains five convolutional layers, an area attention layer, and a fully connected layer. The first convolutional layer is combined with two parallel convolutional layers to extract textures from the horizontal (time axis) and vertical (MFCC axis) respectively. There is a batch normalization layer after each convolutional layer.**

TABLE 2. **Specific settings for the convolutional layers.**

Layer	*Settings*
Conv1a	Kernel size=(10,2),stride=1,in channels=1,out channels=8
Conv1b	Kernel size=(2,8) ,stride=1,in channels=1,out channels=8
Conv2	Kernel size=(3,3) ,stride=1,in channels=16,out channels=32
Conv3	Kernel size=(3,3) ,stride=1,in channels=32,out channels=48
Conv4	Kernel size=(3,3) ,stride=1,in channels=48,out channels=64
Conv5	Kernel size=(3,3) ,stride=1,in channels=64,out channels=80

B. MODEL

FIGURE 2 shows the overall structure of our model, which consists mainly of five convolutional layers and an attention layer. We use the extracted MFCCs as input and treat this input as an image. We use two parallel convolutional layers as the first layer with a kernel size of (10,2) and (2,8) to extract the horizontal (cross-time) and vertical (cross-MFCC) textures. After padding, the 8-channel output of each convolutional layer is concatenated to a 16-channel representation. Then four convolution layers are applied to generate an 80-channel representation and sent to the self-attention layer. Table 2 shows the specific settings for the convolutional layers. After each convolutional layer, a Batch Normalization (BN) layer and an activation function Relu are used, and *conv2* and *conv3* are followed by max-pooling with a kernel size of 2 to reduce the data size.

C. HEAD FUSION

The attention mechanism can be regarded as a soft addressing operation, which uses key-value pairs to represent the content stored in the memory, and the elements are composed of the address (*key*) and the value (*value*). *Query* can match to a key which correspondent *value* is retrieved from the memory according to the degree of correlation between the *query* and the *key*. The *query*, *key*, and *value* are usually first multiplied by a parameter matrix W to obtain Q, K and V. Eq.1 shows the calculation of attention map, where d_k is the dimension of K [24] to prevent the result from being too large. In self-attention, the *query*, *key* and *value* come from a same input X. By calculating self-attention, the model can focus on the connection between different parts of the input. In SER, the distribution of emotion characteristics often crosses a larger scale, and using self-attention in speech emotion recognition improves the accuracy.

$$Q = W_q * query, \quad K = W_k * key, \ V = W_v * value$$
$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{1}$$

Even if the multi-head method generates multiple attention maps with some different weight parameters, the map still has only one focus. Therefore, we proposed the Head Fusion method, hoping to focus on multiple points in the same attention map, so as to better find the relationship between features.

We use X^i_{attn} to represent the i_{th} attention map and calculate X^i_{attn} by using different parameter sets W^i_k, W^i_q, W^i_v, where $i \in (0, n_{head}]$, and each X^i_{attn} is called a 'head'. Different from the general multi-head self-attention, we superimpose heads to obtain an attention map with multiple points of attention–X_{mattn} according to Eq. 2.

$$X_{mattn} = \frac{\sum X^i_{attn}}{n_{head}} \tag{2}$$

FIGURE 3 shows the following working process of head fusion. Then we use a global average pooling (GAP) to generate a feature point X_{fusion} for this map. In the past computer vision research, GAP has been proved to be an effective method [39]. We call this superposition method head fusion. We set hyperparameter n_{head} to represent how many heads being fused in a feature point and n_{size} to represent how many feature points to generate. Finally, we concatenate these points and feed them to the fully connected layer to obtain the final classification result.

V. EXPERIMENT

A. PRE-PROCESSING

1) SELECTING DATA

We use utterance as the unit and randomly divide speech emotion data into an 80% train data set and a 20% test data set. Each utterance has an 80% probability of being selected into the training set and a 20% probability of being selected into the test set. The two are mutually exclusive. Since the division of the data set is random and the number of utterances is large enough, each subset has the same distribution as the original data set.

2) NOISE INJECTION

Since researchers do not intentionally collect speech emotion data with noise, it is a general method for researching

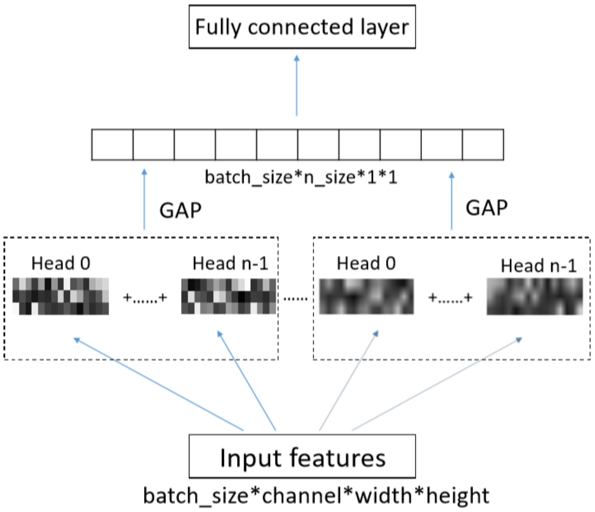


FIGURE 3. Head Fusion 'Head i' represents the i_{th} attention map calculated X^i_{attn} by using different parameter sets W^i_k, W^i_q, W^i_v. Heads are superimposed to obtain an attention map with multiple points of attention. Then a global average pooling (GAP) is applied to generate a feature point X_{fusion} for this map. Finally, these points are concatenated and fed to the fully connected layer to obtain the final classification result. The hyperparameter n_{head} represents how many heads being fused in a feature point and n_{size} represents how many feature points to generate, *i.e.*, a total of $n_{head} * n_{size}$ different attention maps are generated.

TABLE 3. The meaning of the variables in the SNR formula.

Variable	Meaning
N	Number of utterances in speech data set for random selection.
M	Number of utterances in noise data set for random selection.
AMP$_{si}$	Max amplitude of the i_{th} utterance in speech data set.
AMP$_{ni}$	Max amplitude of the i_{th} utterance in noise data set.
AF	Amplitude factor

SER under noisy conditions to synthesize data with noise data set and clean speech emotion data set. We also synthesize noisy data like other related researchers have done [29], [31].

As shown in FIGURE 4, we use different methods to select noise according to the experiments we conducted. In the noise research, we only use the IEMOCAP data set for its naturalness.

We select one piece of noise to mix with each utterance. If the duration of the noise is shorter than the speech utterance, it will be repeatedly concatenated. We use the amplitude factor(AF) to control the intensity of the noise. Before the noise is mixed with the speech data, the amplitude of the noise is first multiplied by AF. The average signal-to-noise ratio(SNR) can be calculated according to Eq.(3) and Table 3 shows the meaning of the variables. When AF is set to 0.1, we evaluated the average SNR and it is approximately 11 dB. When the AF doubles, the SNR decreases by around 6 dB. When AF is set to 0, it means the data is clean without noise.

$$SNR = 20 * \log_{10} \frac{\sum_{i=0}^{N-1} \frac{AMP_{si}}{N}}{\sum_{i=0}^{M-1} \frac{AMP_{ni}}{M} * AF} \tag{3}$$

3) UTTERANCE SLICING

We divide each utterance into several segments with an equal length of two seconds. The remaining segment less than two seconds is removed. The segments preserve the original label and trained individually.

In order to avoid data loss, there is a one-second overlap between two consecutive segments. In the test process, segments from the same utterance are grouped into the same mini-batch, and the result is averaged to obtain the resulting emotion type. In the test data set, the overlap is increased to 1.6 seconds. Experience shows that a large overlap can make the recognition result of utterance more stable in testing.

B. EVALUATION METRICS

Due to the imbalance of the IEMOCAP data set, we mainly use unweighted accuracy(UA) for evaluation, supplemented by weighted accuracy(WA). UA refers to the average accuracy of each emotion category, WA refers to the accuracy of all samples. These two have been broadly employed in SER literature. The two types of accuracy are calculated according to Eq.(4).

$$UA = \frac{\sum_1^k \frac{n_i}{N_i}}{k}, \quad WA = \frac{\sum_1^k n_i}{\sum_1^k N_i} \tag{4}$$

The N_i means the number of utterances in i_{th} class, the n_i means the number of correctly recognized utterances in i_{th} class and the k means the number of classes.

C. EXPERIMENTAL SETUP

We implemented the model with PyTorch, used the cross-entropy loss function, and optimized it with the Adam optimizer. The batch size was set to 32. The initial learning rate was 0.001, weight decay was 1e-6. We trained the model with 50 epochs and evaluated its WA and UA.

Due to the randomness in our experiment (including data set division, model initialization, and training process), we used random seeds to fix the experimental results for optimization and reproduction. To reduce the impact of randomness, we used 5 different random seeds to increase the reliability of the experimental results, *i.e.* we conducted 5 independent experiments, and finally calculated the average of the accuracy as the final experimental result.

D. EXPERIMENTAL RESULTS

1) EXPERIMENTS ON THE CLEAN DATA SET

In the experiment, the accuracy is improved compared to the self-attention model without head fusion, and the value of n_{head} has a significant effect on the experimental results. We find that as n_{head} increases, the accuracy increases gradually and reaches a maximum at a certain point. Then, if n_{head} continues to increase, the accuracy will gradually decrease. We believe that the reason for this decline is that the value of n_{head} is too high, causing some attention points in the map to be placed in too much detail, which will reduce the effect of attention instead.

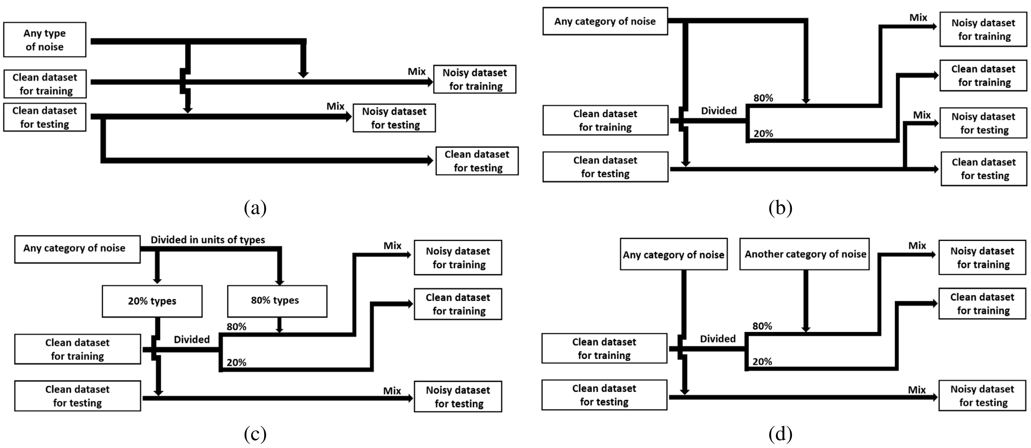
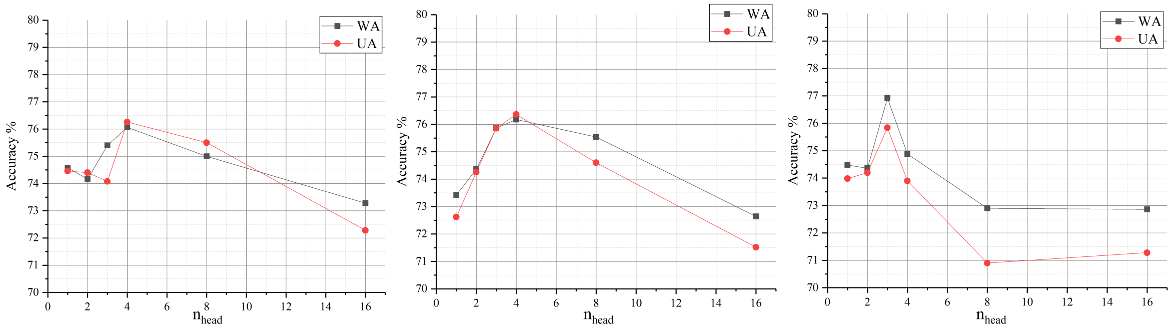

Any type of noise
Clean dataset for training
Clean dataset for testing
Mix
Noisy dataset for training
Mix
Noisy dataset for testing
Clean dataset for testing
(a)
Any category of noise
Clean dataset for training
Clean dataset for testing
Divided
80%
20%
Mix
Noisy dataset for training
Clean dataset for training
Mix
Noisy dataset for testing
Clean dataset for testing
(b)
Any category of noise
Divided in units of types
20% types
80% types
Clean dataset for training
Divided
80%
20%
Clean dataset for testing
Mix
Noisy dataset for training
Clean dataset for training
Mix
Noisy dataset for testing
(c)
Any category of noise
Another category of noise
Clean dataset for training
Divided
80%
20%
Clean dataset for testing
Mix
Noisy dataset for training
Clean dataset for training
Mix
Noisy dataset for testing
(d)

FIGURE 4. Different experiment settings on various noise injection strategies. (a): we randomly select a snippet from a type(*e.g.* dog) for each target speech utterance to generate the noisy train and test data set. We test the model on both noisy (same type as training) and clean test data. (b): we split train data into 80% (mixing with noise) and 20% (keeping clean). We use all types of noise from a specific category (*e.g.* D-S). By splitting the training data we would mimic real-life scenarios where clean data is insufficient. We call 20% a unit, *i.e.* 1 unit for clean data and 4 for noisy data. (c): the training data is again split 80%/20%. We first choose a category and the noises in the chosen category (*e.g.* D-S) are further randomly split into 80%/20% by type, where 80% types are added to the 80% clean train and 20% types to the 20% clean test to evaluate. (d): similar to (c) in data split, with the difference of adding different categories of noises instead of different types within the same category as in (c), to evaluate the inter-category impact.



(a) n_{size}=32 (b) n_{size}=64 (c) n_{size}=128

FIGURE 5. Modifying the n_{head} when setting the n_{size} to 32, 64 and 128. It can be seen that when n_{size} is set to 32 or 64, the highest accuracy is obtained when n_{head} is set to 4, and when n_{size} is set to 128, the highest accuracy is obtained when n_{head} is set to 3. As n_{head} increases, the accuracy first increases and then decreases.

FIGURE 5 shows the change in accuracy caused when modifying the n_{head} when n_{size} is set to 32, 64, and 128. When n_{size} is set to 32 or 64, we obtain the highest accuracy when n_{head} is set to 4, and when n_{size} is set to 128, we obtain the highest accuracy when n_{head} is set to 3. We set n_{head} to 4 in our final model.

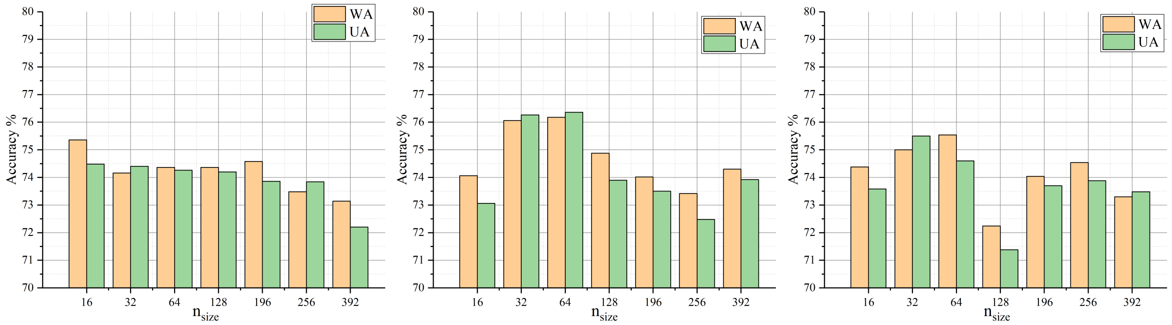
As shown in FIGURE 6, we also tested the effect of different n_{size}s on accuracy with n_{head} set to 2,4 and 8. When n_{head} is set to 2, we obtain the highest accuracy when n_{size} is set to 16, when n_{head} is set to 4, we obtain the highest accuracy when n_{size} is set to 64, and when n_{head} is set to 8, we obtain the highest accuracy when n_{size} is set to 32. We set n_{size} to 64 in our final model.

According to the experimental results, we believe that n_{head} has a significant impact on accuracy, and n_{size} has an impact on accuracy but not significant enough. When we fix the value of n_{size} and change n_{head} slightly, we can easily obtain the maximum accuracy of a model (in our model, this value is around 76%). However, for n_{size}, fixing the value of n_{head} and changing n_{size} has an effect on the accuracy of the model. In that, we recommend that choose a suitable n_{size} (such as 32, 64, *etc.*) and modify n_{head} to obtain the maximum accuracy when using the head fusion method, instead of focusing on adjusting n_{size}.

FIGURE 7(a) shows the confusion matrix of our experiment on the IEMOCAP data set.

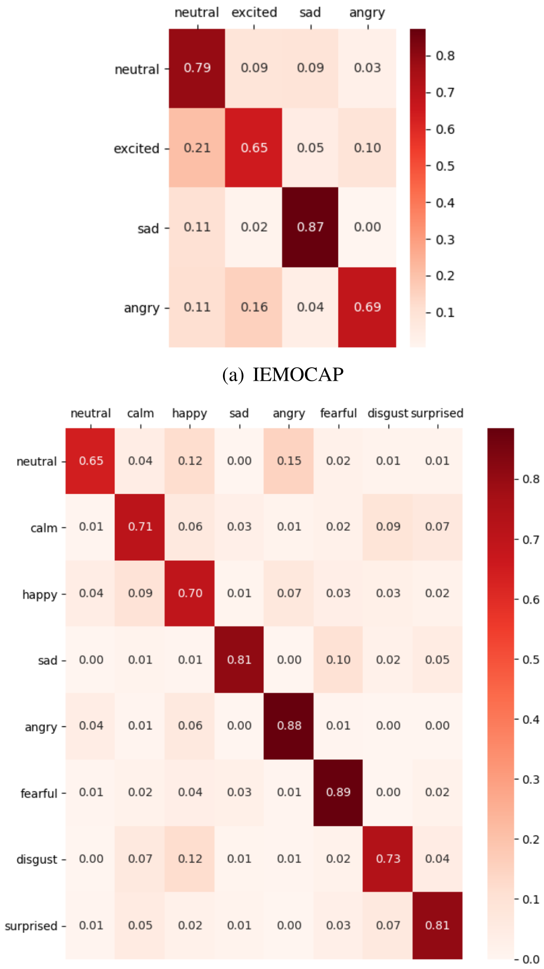
We provide a comparison of the accuracy of previous research and our model in Table 4, all of the experiments used the improvised part of IEMOCAP as data set. It can be seen that our model is state-of-the-art, especially the UA has been greatly improved (about 6%).

To compare with the ACNN model based on conventional self-attention [20], we tested our best model on the scripted part and the full IEMOCAP data set. The result is shown in Table 5.



(a) n_{head}=2 (b) n_{head}=4 (c) n_{head}=8

FIGURE 6. Modifying the n_{size} when setting the n_{head} to 2, 4 and 8. It can be seen that when n_{head} is set to 2, the highest accuracy is obtained when n_{size} is set to 16, when n_{head} is set to 4, the highest accuracy is obtained when n_{size} is set to 64, and when n_{head} is set to 8, the highest accuracy is obtained when n_{size} is set to 32.



(a) IEMOCAP

(b) RAVDESS

FIGURE 7. Confusion matrix The vertical axis represents the label of a sample. The horizontal axis represents the label predicted by the model. The data in the figure represents the probability that the model predicts the vertical axis label as the horizontal axis label.

We also conducted an ablation experiments to verify the effect of the attention layer and compare the effects of the number of intermediate convolution layers. The result is shown in Table 6.

TABLE 4. Comparison with existing SER results on the IEMOCAP data set.

Method	*WA(%)*	*UA(%)*	*Year*
Attention pooling (P. Li *et al.*) [21]	71.75	68.06	2018
CTC + Attention (Z. Zhao *et al.*) [27]	67.00	69.00	2019
Self attention (L. Tarantino *et al.*) [20]	70.17	70.85	2019
BiGRU (Y. Xu *et al.*) [4]	66.60	70.50	2020
Multitask learning + Attention (A. Nediyanchath *et al.*) [6]	76.40	70.10	2020
Head fusion (Ours)	**76.18**	**76.36**	2020

TABLE 5. Comparison with the ACNN model based on conventional self-attention.

Method	*WA(%)*	*UA(%)*
self-attention (L. Tarantino *et al.* [20]) (improvised)	70.17	70.85
self-attention (L. Tarantino *et al.* [20]) (scripted)	64.59	50.12
self-attention (L. Tarantino *et al.* [20]) (full)	68.10	63.80
Our model (improvised)	76.18	76.36
Our model (scripted)	65.90	63.92
Our model (full)	67.28	67.94

TABLE 6. Ablation experiments.

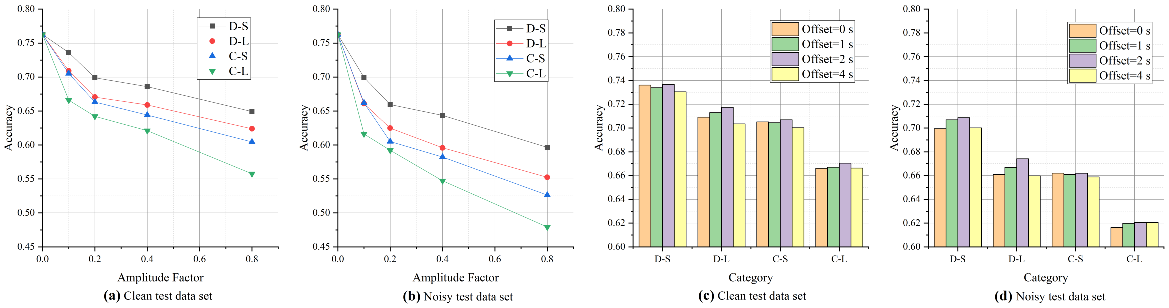
Method	*WA(%)*	*UA(%)*
1 convolutional layer + head fusion	63.94	60.14
2 convolutional layers + head fusion	73.94	74.36
3 convolutional layers + head fusion	74.68	74.66
4 convolutional layers + head fusion	76.18	76.36
4 convolutional layers without head fusion	68.7	66.9

To verify the generality of our proposed method, we also validate our model on the RAVDESS dataset. FIGURE 7(b) shows the confusion matrix of our experiment on the RAVDESS data set. It can be seen that the accuracy of the neutral class is relatively low. We suggest that it is because the number of samples in the neutral class is half of that in other classes in the RAVDESS data set, resulting in the model not being able to learn the characteristics of the neutral class well.

Table 7 shows a comparison with previous studies using the RAVDESS data set. It can be seen that our method is state-of-the-art.

2) EXPERIMENTS ON THE NOISY DATA SET

In this part, we mainly research the impact of different types of noise in SER, especially the effect of adding additional



(a) Clean test data set **(b)** Noisy test data set **(c)** Clean test data set **(d)** Noisy test data set

FIGURE 8. The training scheme complies with FIGURE 4(a). The average UA of the four categories is shown. The conclusion is adding D-S noise affects accuracy the least while C-L the most, a gap of 7% on clean test data and 8.3% on the noisy test (AF is set to 0.1). Both (a) and (b) shows increasing AF affects the accuracy significantly across all noise categories but the relative performance trend retains. Both (c) and (d) suggest a trivial difference in the choice of noise offset shifting.

TABLE 7. Accuracy comparison with existing SER results on the RAVDESS data set.

Method	*WA(%)*	*UA(%)*	*Year*
GResNet+Multi-task (Y. Zeng *et al.*) [40]	64.5	64.6	2019
BLSTM+Capsule routing (M. Jalal *et al.*) [41]	/	69.40	2019
DCNN (D. Issa*et al.*) [42]	/	71.6	2020
Multimodal fine-grained learning (H. Li *et al.*) [43]	77.0	74.7	2020
Head Fusion (Ours)	**77.8**	**77.4**	2020

noise data for training. On one hand, we test the robustness of our method and compare it with previous works. This makes sense since the speech is always mixed with some noise in real life. On the other hand, We suggest that this helps alleviate the problem of insufficient high-quality data in SER research. We hope to find out the categories of noise that have less impact on training, and add natural speech with these categories of noise to the training. This not only reduces the cost of professional recording studios and actors, and the speech obtained is more natural.

a: THE INFLUENCE OF EACH CATEGORY OF NOISE ON SER

In this section, we tested the impacts of the 50 types of noise on SER. The method of noise selection is shown in FIGURE 4(a). We trained the model on the noisy data set and tested it on both the clean data set and the noisy data set respectively.

We conducted a T-test on the UA and found that the UA of the model trained with D-S noise was significantly higher($p<0.001$) than that of all other categories, and the UA of the model trained with C-L noise was significantly lower($p<0.001$). The margin of average UA between them is more than 7%. The margin between using D-L or C-S noise is not significant($p>0.05$) and the UA is close. As shown in FIGURE 8(a) and 8(b), we calculated the average UA of the categories and conducted the experiments under different AF conditions. It can be considered that data mixed with C-L is not a good choice for training.

In addition, as shown in FIGURE 8(c) and 8(d), we also shifted the time axis of the noise with different offsets by 1, 2, and 4 seconds. The result shows that injecting noises at varied time offset has a negligible influence on the UA.

b: USING ADDITIONAL NOISY DATA FOR TRAINING

In this section, we simulate the case of adding noisy data when high-quality data is insufficient.

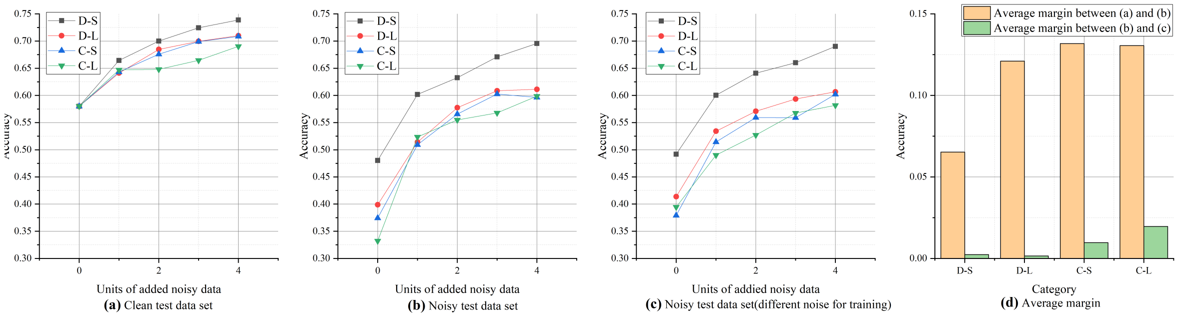
First, we selected noise according to FIGURE 4(b). As shown in FIGURE 9(a) and (b), adding noisy data for training can effectively improve UA. Respectively on the clean and noisy test data set, the D-S data achieved the highest UA(73.8% and 69.5%), and the C-L data achieved the lowest UA(69% and 59.8%). In addition, in the noisy test data set, compared to training without noisy data, even adding 1 unit of the C-L data can greatly improve the UA from 33.2% to 52.3%.

Then, we selected noise in the method of FIGURE 4(c) to verify whether the models obtained by training with certain types of noise will help de-noise data mixed with other noise types within the same category.

As shown in FIGURE 9(c), results similar to FIGURE 9(b) can still be achieved. Compared with using the same noise type for training and testing, the UA has decreased but not great.

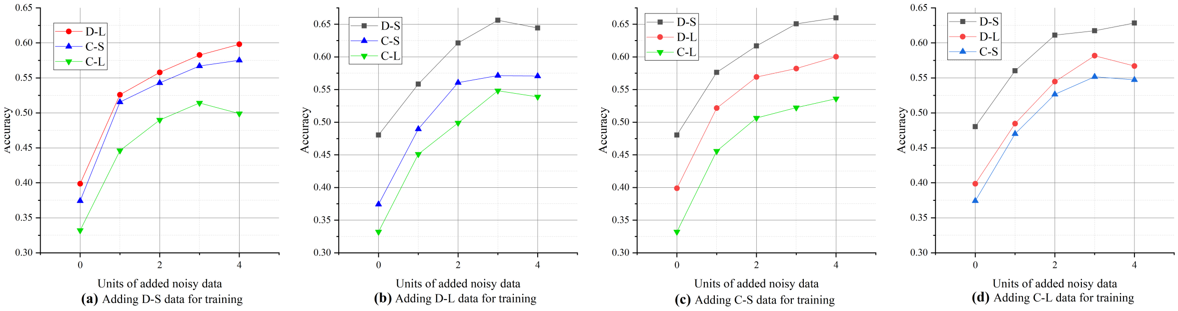
As shown in FIGURE 9(d), the average margin of the category *Discrete* (D-S or D-L) is less than 1%, and that of category *Continuous*(C-S or C-L) is less than 2%. It is caused by the model trained within one noise category could adapt to other noise types within the same noise category. In other words, within the same category of noises, one type of noise can be substituted by others for training without degrading the test accuracy.

Last, we selected noise in the method of FIGURE 4(d). The only difference with FIGURE 9(c) would be the noise for testing from a different category. As shown in FIGURE 10, the UA also increases with the noisy data quantity increasing, but compared to FIGURE 9, the UA further decreased by over 2% in all cases, especially those involving category C-L, which dropped by more than 4%. In addition, when the quantity of noisy data added to 4 units, the UA of some categories decreased, which could be caused by overfitting. Based on the experimental results, we suggest that it is a



(a) Clean test data set **(b)** Noisy test data set **(c)** Noisy test data set(different noise for training) **(d)** Average margin

FIGURE 9. (a) and (b) adopts the training scheme of FIGURE 4(b) which shows the UA increases with the quantity of noisy train data added to clean train data (AF = 0.1). The UA increases by 15.9% by adding 4 units of D-S noise as in (b). Both the clean (see (a)) and noisy test data (see (b)) shows a similar upward trend. (c) applies the training scheme of FIGURE 4(c) with different types of noises within the same noise category. It shows a negligible difference of UA to (b). In (d), the orange bar represents the average difference of UA between (a) and (b), and the green bar represents that between (b) and (c). It can be seen that the green bar of the D-S and the D-L is less than 1% and that of the C-S and C-L is less than 2%, suggesting the model is robust enough to deal with different types of noises within the same category due to the similarity of noises.



(a) Adding D-S data for training **(b)** Adding D-L data for training **(c)** Adding C-S data for training **(d)** Adding C-L data for training

FIGURE 10. The training scheme adopts FIGURE 4(d). AF is set to 0.1. Similar to FIGURE 9(c), the train data includes one unit of clean data and up to four units of noisy data, with the difference that the noises added to training and testing are from different categories, *e.g.* in FIGURE 10(a), we use D-S for training and use D-L, C-S and C-L for testing. It shows that in general, the UA increases with the quantity of noisy data added on top of the clean train data. However, when the quantity of noisy data added from 3 units to 4 units, the UA of some categories decreased. Our hypothesis is when a small amount of noise train data is added, the model learns to de-noise. But when more is added, the model overfits to the category of noise added.

good choice to use a noise data quantity that is three times the clean data quantity. We also modified the AF to repeat the above experiments but the model differences are similar across several AF choices (0.1,0.2,0.4 and 0.8).

C: COMPARISON WITH PREVIEWS WORK

In this section, to verify the robustness of our model, we compare our experimental results with the bio-modal system and the single-audio system without designed facial expression recognition designed by Y. Huang *et al.* in 2019 [31], since we used similar experimental metrics and data set. To make the results more universal, we do not distinguish the types of noise, and randomly select noise from all noises to generate noisy data for training. Huang *et al.* carried out experiments under noisy condition with different signal-to-noise ratios (SNR). Since our noise is randomly selected, we calculated the average SNR under different AF for comparison according to Eq.(3).

As shown in Table 8, on the clean data set, our model achieves higher UA than the single-audio system without facial expression recognition. When the SNR is about 15 (AF = 0.1), our model achieves UA similar UA to the bio-modal system, which is much higher than the single-audio system. When AF = 0.5 (SNR ≈ 1.4), our model achieves lower UA than the bio-modal system under the condition of SNR = 5 but higher than the bio-modal system under the condition of SNR = 0. The UA of our model is also much higher than that of the single-audio system when SNR = 5. It can be seen that with the noise increasing, the accuracy of our model decreases more slowly. which shows that our model has better robustness.

TABLE 8. Comparison with previews work.

Method	UA	SNR	AF
Our model (speech)	72.26	clean	0
bio-modal system (speech + expression)	80	clean	/
single-audio system (speech)	70	clean	/
Our model (speech)	59.42	15.3	0.1
bio-modal system (speech + expression)	60	15	/
single-audio system (speech)	54	15	/
Our model (speech)	48.04	1.4	0.5
bio-modal system (speech + expression)	51	5	/
single-audio system (speech)	37	5	/
bio-modal system (speech + expression)	46	0	/
single-audio system (speech)	30	0	/

VI. CONCLUSION

In this paper, we propose an improved mechanism head fusion for SER based on multi-head self-attention and verify the role of its parameters. We implemented an ACNN model and conducted experiments on the IEMOCAP and RAVDESS data set, achieving the state-of-the-art accuracy.

On the other hand, we found that adding noisy data is a viable solution when high-quality data is insufficient in SER research. It also improves the robustness of the model. Adding data with D-S noise achieves the highest accuracy and adding data with C-L noise achieves the lowest accuracy. We hypothesis that D-S noise is similar to human pronunciation rules, and is easier to be found and eliminated by models that recognize human voices.

When we train our model with data mixed with the noise of one category, it achieves good robustness and extends well to other types of noise within the same noise category. It is recommended to use noisy data three times over the clean data to prevent overfitting to a certain category. Besides, we found that the intensity of noise affects the accuracy significantly and the relative performance trend retains, but the time offset has a negligible influence on the accuracy.

In the future, we will continue to work along the lines by extending various de-noising and data augmentation strategies based on our established conclusion.

ACKNOWLEDGMENT

The authors would like to acknowledge Prof. Samee U.Khan for his contribution on advising and discussing to improve the quality of the manuscript.

REFERENCES

[1] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, 2014, pp. 223–227.

[2] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[3] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. Meng, "Speech emotion recognition using capsule networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6695–6699.

[4] Y. Xu, H. Xu, and J. Zou, "HGFM: A hierarchical grained and feature model for acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 6499–6503.

[5] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 3227–3231.

[6] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 7179–7183.

[7] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *Proc. 10th Annu. Comput. Commun. Workshop Conf. (CCWC)*. Las Vegas, NV, USA: IEEE, Jan. 2020, pp. 1058–1064.

[8] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Comput. Hum. Behav.*, vol. 93, pp. 309–317, Apr. 2019.

[9] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.

[10] J. Yang, F. Zhang, B. Chen, and S. U. Khan, "Facial expression recognition based on facial action unit," in *Proc. 10th Int. Green Sustain. Comput. Conf. (IGSC)*. Alexandria, VA, USA: IEEE, Oct. 2019, pp. 1–6.

[11] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 20–29.

[12] L.-S.-A. Low, M. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in Adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011.

[13] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, vol. 1. Sanya, China: IEEE, Oct. 2010, pp. 537–541.

[14] W. J. Yoon, Y. H. Cho, and K. S. Park, "A study of speech emotion recognition and its application to mobile services," in *Proc. Int. Conf. Ubiquitous Intell. Comput.* Berlin, Germany: Springer, Jul. 2007, pp. 758–766.

[15] G. F. Arnold and J. O'Connor, *Intonation of Colloquial English*. London, U.K.: Longman, 1973.

[16] P. Ekman and H. Oster, "Facial expressions of emotion," *Annu. Rev. Psychol.*, vol. 30, no. 1, pp. 527–554, 1979.

[17] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[18] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, no. 3, pp. 273–294, Sep. 1977.

[19] R. Altrov and H. Pajupuu, "The influence of language and culture on the understanding of vocal emotions," *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri., J. Estonian Finno-Ugric Linguistics*, vol. 6, no. 3, pp. 11–48, Dec. 2015.

[20] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2578–2582.

[21] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 3087–3091.

[22] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2. Hong Kong: IEEE, 2003, pp. 1–2.

[23] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[25] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," 2019, *arXiv:1904.13377*. [Online]. Available: http://arxiv.org/abs/1904.13377

[26] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," 2019, *arXiv:1906.09890*. [Online]. Available: http://arxiv.org/abs/1906.09890

[27] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 206–210.

[28] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Brighton, U.K.: IEEE, May 2019, pp. 2822–2826.

[29] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Robust front-end processing for emotion recognition in noisy speech," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*. Taiwan, China: IEEE, Nov. 2018, pp. 324–328.

[30] Y. Huang, K. Tian, A. Wu, and G. Zhang, "Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 5, pp. 1787–1798, May 2019.

[31] Y. Huang, J. Xiao, K. Tian, A. Wu, and G. Zhang, "Research on robustness of emotion recognition under environmental noise conditions," *IEEE Access*, vol. 7, pp. 142009–142021, 2019.

[32] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Proc. 24th ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, Oct. 2016, pp. 571–575.

[33] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," 2019, *arXiv:1905.02921*. [Online]. Available: http://arxiv.org/abs/1905.02921

[34] 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 database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, Dec. 2008.

[35] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Brighton, U.K.: IEEE, May 2019, pp. 7390–7394.

[36] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[37] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*. Brisbane, QLD, Australia: ACM, 2015, pp. 1015–1018.

[38] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, vol. 8, Jul. 2015, pp. 18–25.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[40] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019.

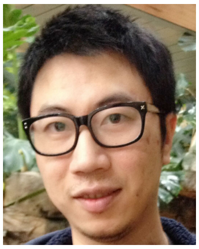
[41] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1701–1705.

[42] D. Issa, M. F, Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894.

[43] H. Li, W. Ding, Z. Wu, and Z. Liu, "Learning fine-grained multimodal alignment for speech emotion recognition," 2020, *arXiv:2010.12733*. [Online]. Available: http://arxiv.org/abs/2010.12733

MINGKE XU is currently pursuing the master's degree with the Nanjing Tech University, Nanjing, China. His main research interest includes speech emotion recognition.

FAN ZHANG received the Ph.D. degree in control science and computer engineering from Tsinghua University, China, in 2012. His previous work primarily focuses on managing and orchestrating virtualized resource provisioning for fluctuating workloads in elastic cloud platforms. His main research interests include cloud computing, big-data processing, artificial intelligence, characterizing, tuning and optimizing big-data applications, such as Hadoop MapReduce, Giraph Pregel, and Graphlab.



WEI ZHANG received the degree from CMU, in 2014. He is currently a Research Scientist with IBM T.J. Watson Research Center. He is broadly interested in artificial neural networks and natural language processing. His recent interest is in explainable NLP, human-AI and AI-AI communication in dialogues and language games, and NLP applications to high-stake domains, such as finance.