IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 26, 2020, accepted December 6, 2020, date of publication December 10, 2020,
date of current version December 24, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3043894*

# Exploration of an Independent Training Framework for Speech Emotion Recognition

**SHUNMING ZHONG[1,2], BAOXIAN YU[1,2,3], AND HAN ZHANG[1,2,3], (Member, IEEE)**

[1]School of Physics and Telecommunication Engineering, South China Normal University (SCNU), Guangzhou 510006, China
[2]Guangdong Provincial Engineering Technology Research Center of Cardiovascular Individual Medicine & Big Data, SCNU, Guangzhou 510006, China
[3]SCNU Qingyuan Institute of Science and Technology Innovation Company Ltd., Qingyuan 511517, China

Corresponding authors: Han Zhang (zhanghan@scnu.edu.cn) and Baoxian Yu (yubx@m.scnu.edu.cn)

This work was supported in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515011940, in part by the Science and Technology Program of Guangzhou under Grant 2019050001 and Grant 202002030353, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2017B030308009, in part by the Special Project for Youth Top-Notch Scholars of Guangdong Province under Grant 2016TQ03X100, and in part by the Scientific Research Cultivation Project for Young Scholars of South China Normal University under Grant 19KJ16.

**ABSTRACT** Speech emotion recognition (SER) plays an indispensable role in human-computer interaction tasks, where the ultimate performance is determined by features, such as empirically learned features (ELFs) and automatically learned features (ALFs). Although the fusion of both ELFs and ALFs can complement some new features for SER, the fused training within one softmax layer is inappropriate due to the different performance of using either ELFs or ALFs for emotion recognition. Based on this consideration, this paper proposes an independent training framework that can fully enjoy the complementary advantages of human knowledge and powerful learning ability of deep learning models. Specifically, we first feed Mel frequency cepstral coefficient and openSMILE features respectively into a pair of independent models, which are composed of an attention-based convolution long short-term memory neural network and a fully connected neural network. We then design a feedback mechanism for each model to extract ALFs and ELFs independently, where hard example mining and re-training with a hard example loss are applied to focus the feature extraction on hard examples during training. Finally, a classifier is adopted to distinguish emotion by using both the independent features of ALFs and ELFs. Based on extensive experiments on three public speech emotion datasets (IEMOCAP, EMODB, and CASIA), we show that the proposed independent training framework outperforms the conventional feature fusion methods.

**INDEX TERMS** Data imbalance, hard example, feature fusion, independent training, speech emotion recognition.

## I. INTRODUCTION

Speech is the most common and natural communication medium for human beings. In addition to the linguistic information, speech signals also contain some useful paralinguistic information, such as gender, age and emotion [1], [2]. In recent ten years, speech emotion recognition (SER), as one of the most challenging techniques in the field of human machine interaction (HMI) [1], [3], has been widely applied to education, health, smart home and many other artificial intelligence scenarios [4].

With the development of deep learning technology, the raw signals and their spectrum can be directly used as the input of models for many classification tasks, including the field of SER [5]–[7]. Such model is referred to as ''end-to-end'' system since it requires no pre-process on the raw signals, and can yield satisfactory classification performance thanks to the power ability of the specifically designed deep model.

In order to alleviate the computational cost of feature extraction using raw signals, some pre-process is considered by the authors in [8], where specific human-empowered features are generated from the raw signals and then fed into classifiers for emotion recognition. In general, discriminative feature generation is the most important step for SER task, since the features extracted from speech signals contain effective information of emotional states [9]–[11]. Previous works demonstrated that empirically learned features (ELFs), such as prosodic, vocal tract, pitch, and voice quality feature *et. al.* can characterize emotion states clearly [12]–[14], and accordingly, a variety of emotion feature sets have been used for SER task, e.g., INTERSPEECH 2013 [15], AVEC-2013 [16] and GeMAPs [17]. However, the performance of such features might be limited by the fact that they are hand-crafted, that is, designed by human experimenters, and might therefore not be optimal to characterize vocal emotions.

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

As an alternative, automatically learned features (ALFs) provides a promising and an outstanding solution for SER thanks to the development of deep learning techniques [18]. In comparison with ELFs, ALFs extracted based on MFCC are more effective to capture high-level features by taking temporal-spatial relation of feature map into consideration [19]–[21]. Typically, the convolutional neural network (CNN) model or the long short-term memory (LSTM) model is used in ALF-aided emotion recognition [22]–[24]. To capture more useful emotional feature representations, Mao *et al.* [25], proposed a two-stage CNN, which consists of sparse auto-encoder and salient discriminative feature analysis, to learn affect-salient discriminative features. Zhao *et al.* [26] designed a merged deep model, where both 1D and 2D CNN were applied to learn high-level features from raw audio clips and Log-Mel spectrograms, respectively. In order to characterize the dynamic characteristic of speech signal, 3D-based model is considered, where Chen *et al.* [27] introduced 3D-based attention convolutional recurrent neural networks (CRNN). Meng *et al.* [19] developed a dilated CNN, which selected 3D Log-Mel spectrograms as input. Peng *et al.* [28] proposed 3D convolutions and attention-based sliding recurrent networks to fully utilize the auditory and attention mechanism of human. In addition, some parallel deep model works have also been explored. Jiang *et al.* [29] employed Log-Mel spectrograms for CNN and frame-level feature for LSTM. Zhao *et al.* [30] applied spectrogram as the input of Bi-LSTM and CNN separately to yield different highly-abstract feature representations. Although ALFs of deep models can generally capture effective emotional information, the performance of feature fusion between ALFs and ELFs is not clearly demonstrated by the above mentioned works, but could have the potential of improving the accuracy of SER.

In recent two years, some pioneer works focused on the fusion of ELFs and ALFs for SER. The idea behind the fusion process is to utilize the complementary advantages between ALFs and ELFs. In [31]–[34], the authors introduced the fusion of features with both ALFs and ELFs, and then extended the idea of fusion by taking the advantage of spectrogram and openSMILE features with CNN and deep neural network (DNN), respectively. In comparison with SER tasks using one kind of feature, the aforementioned methods employing both ELFs and ALFs can improve SER performance. However, the effect of fusion at the feature level is neglected by the existing contributions. To elaborate a little further, the system employing ELFs and ALFs performs differently when tracking identical tasks, and thus requires different feedback of loss functions to optimize the corresponding model. In other words, the fusion of ELFs and ALFs requires independent models of different feedback to achieve the global optimization, which is ignored by the existing contributions.

In order to take full advantages of both ELFs and ALFs, this paper proposes an independent training framework. Unlike the conventional fusion works [32], [33], the training process of both ELFs and ALFs are completely independent prior to the fusion of features. To be specific, with the independent feedback of loss functions, attention-based convolution long short-term memory network (ACLN) is considered to extract high-level ALFs with spatial-temporal characteristics of emotions, while the fully connected neural network (FCN) is employed to characterize ELFs of granularity characteristics. Finally, we integrate the output of independent model (ALFs and ELFs) into support vector machine (SVM) for emotion recognition directly, when the loss of independent model tends to steady.

With respect to the independent feedback propagation of both ACLN and FCN, we develop a hard example innovated error feedback mechanism. Specifically, we focus on the hard examples during the error feedback process, and develop a hard example (HE) loss to increase the impact of the hard examples on discriminating emotion features. Basically, hard examples belong to the samples of the emotional state that are of a relative high misclassification rate. Typically, in a known emotion dataset, such as the IEMOCAP database, hard examples are concentrated between samples of neutral state and those of happy. In comparison with the problem of data-imbalance, the effect of hard example on SER is ignored by the pioneer works [35], [36]. In [37], Tripathi *et al.* first considered the effect of hard examples in the training process, and developed focal loss based error feedback according to the classification error rate in a residual CRNN model. Although the focal loss takes hard examples into consideration during the error feedback, it is unable to increase the ratio of hard examples over the whole training dataset, leading to limited performance improvement to the effectiveness of feature extraction and emotion discrimination. Moreover, the modification of hyper-parameters in focal loss is necessitated in the training process [37].

In order to address the above issue, this paper presents an independent error feedback mechanism for both ACLN and FCN, where we specially emphasize the impact of hard examples on the feature extraction, and propose a hard example innovated loss in the error feedback propagation. To be specific, we define the hard-index by taking classification error rate into consideration, and then apply hard example mining process to identify hard examples in the whole training set. Based on the so designed hard-index, we develop a HE loss and re-training operation to address the class imbalance problem while emphasizing the contribution of hard examples to the total loss.

To sum up, the contributions of this paper can be summarized as two-folds.

1) We demonstrate through numerical results that the independent training prior to the fusion of features is able to circumvent the limitation of the conventional feature fusion at the feature level. To be specific, we propose to design an independent framework for feature extraction prior to the fusion of features. With independent training and error feedback, the extracted features can better represent the emotions with different levels of

loss, and thus, have the potential of improving the SER performance.

2) We propose a hard example aided re-training method in the process of independent feature extraction. Specifically, we develop the hard-index for hard example mining, and proposed a HE loss by taking advantages of the hard-index and the distributions of emotion classes into consideration. With the designed HE loss and re-training method, the problems of insufficient learning on hard examples and the imbalanced data distribution can be well tackled.

The rest of this paper is organized as follows. Section II describes the background theory and limitations to the conventional training and feedback propagation for SER. We propose the independent training framework in Section III, including the independent feature extraction and feedback mechanism. The experimental results and analysis are presented in Section IV, and Section V concludes this paper.

## II. THEORETICAL BACKGROUND

In comparison with tremendous works of handcrafted feature generation, deep learning model has the ability of extracting features automatically. As the benchmark in the field of SER, the authors in [29], [30] considered CNN and LSTM directly on spectrograms to extract deep acoustic features for emotion recognition. Although these deep models yield good results for many speech processing tasks, ALFs extracted by a specific model like CNN is generally not sufficient due to the lack of human knowledge. It has been pointed out by [32] that ALFs and ELFs can characterize the emotions from different aspects, and some handcrafted empirical features (e.g., F0, MFCC features, energy, and voice probability) are also key issues for distinguishing emotion states. Owing to the distinct merits of both ALFs and ELFs to SER tasks, the authors in [31] demonstrated that the fusion of ALFs and ELFs at the feature level can complementary each other, and thereby, improve the SER performance in comparison with either of them [38], [39]. As an extension work, the authors in [33] presented a CNN-DNN model for feature extraction of both spectrogram and auditory based features, and incorporate both ALFs and ELFs into one feature vector for SER. Unfortunately, the aforementioned fusion works ignored the different performance of both ALFs and ELFs when tracking identical tasks. To be specific, the training of either ALFs or ELFs has the different level of loss, and the fusion of two types of feature prior to the same softmax layer is unable to optimize the corresponding model with different loss. To elaborate a little further, the feature fusion scheme in [31], [33] may fail to fully explore the complementary advantages of both ALFs and ELFs, but has not been addressed by the existing literatures.

As an another key issue in SER, hard examples within the training database remarkably limits the ultimate performance. Basically, hard examples are mainly from the misclassification between the neutral state and the other ones (i.e., happy, angry, etc.). The potential reason is that emotions contained in most of our daily speech is a combination of both verbal and non-verbal channels, i.e., tones and energy of the speech, facial expressions, torso postures, etc. [40]. That means, useful emotion information in some specific spoken utterance are insufficient. These cases are recognized as hard examples, and lead to a high misclassification rate, especially between two adjacent emotional categories. To address the above issue, the authors in [37] applied a feedback mechanism using a specific loss function, namely, focal loss, which is dynamically scaled by cross-entropy (CE) loss. In fact, the principle behind the mechanism of focal loss is to down-weight the contribution of easy examples during training, and focus the model on hard examples. In practice, however, the proportion of hard examples is much smaller than that of easy examples in most training sets, which may easily result in loss gradient dominant of easy examples [41]. It is thus difficult for any deep model to fully characterize the emotional features based on a limited number of hard examples. In other word, it is expected that increasing the proportion of hard examples (reducing the amount of easy examples).

## III. PROPOSED INDEPENDENT TRAINING FRAMEWORK

In this section, we present an independent training framework. As shown in Fig. 1, we feed both MFCC and 384-dimension openSMILE features into ACLN and FCN, respectively. Unlike the pioneer work of fusion at the feature level [33], the two models have independent feedbacks with specifically designed loss functions prior to the fusion of features. The motivation behind the independent design of feature extraction is that MFCC and openSMILE features have different discriminative capabilities for emotion recognition. In other words, the fusion of features with independent models (of independent feedback) enables a better representation of emotion features in comparison with the conventional feature fusion methods. As regards the independent feedback procedure, we develop a new feedback mechanism by taking advantage of hard examples as well as the so designed loss function. The motivation behind the proposed independent feedback mechanism is to emphasize the impact of hard examples during the training procedure. Accordingly, the loss function is modified to enable an effective way to focus the independent model on hard examples while reducing the contribution of easy examples.

### A. DATA PREPROCESSING

For preprocessing of MFCC, we first apply pre-emphasis to improve the high frequency of speech signals, and then use a 24-ms sliding Hamming window with 12-ms overlapping for speech segmentation. Next, we apply short-time Fourier transform of 256 points to obtain the power spectrum w.r.t. the spectrogram of each segments, which is then fed into Mel-scale triangular filter bank to attain Mel energy. Finally, logarithmic function and discrete cosine transform are employed to obtain cepstral coefficient, by which MFCC is generated accordingly.

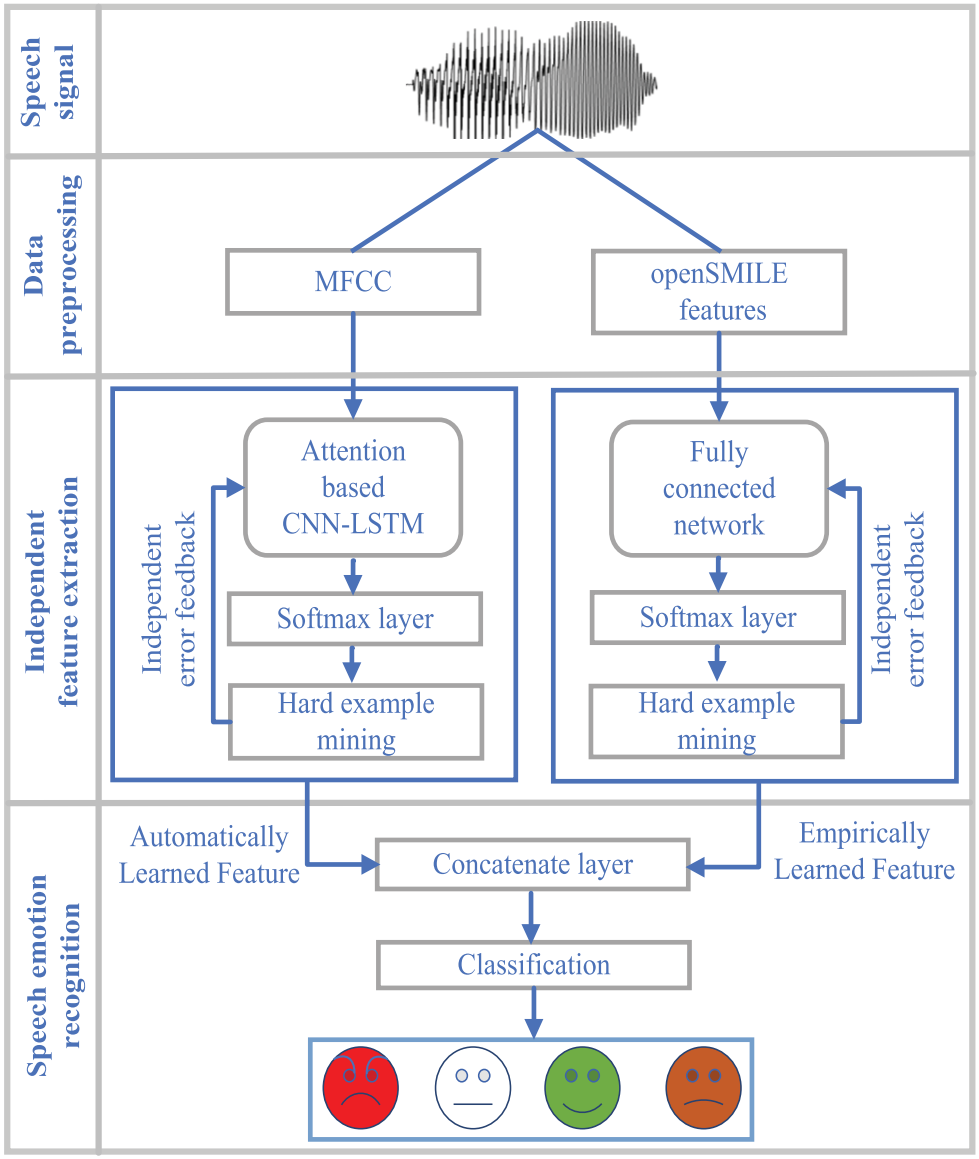



**FIGURE 1.** The framework of the proposed independent training mechanism with hard example loss.

As an alternative, we use openSMILE toolkit [42] to extract openSMILE features of 384 dimensions proposed in [43]. To be specific, features include F0, voice probability, zero crossing rate, energy, MFCC features, etc, totally 32 low-level descriptors. Then, 12 statistical functions are used to generate high-level features, i.e., max, min, mean, range, max-position, min-position, standard deviation, skewness, etc.

### B. INDEPENDENT FEATURE EXTRACTION

The process of independent feature extraction consists of two independent channels of MFCC and openSMILE features, where each channel has its own softmax layer followed by an independent error feedback mechanism.

Since MFCC is efficient to characterize the properties of human auditory perception, we adopt a deep model of ACLN to generate more comprehensive ALFs. As regards the design of ACLN, CNN is first adopted to extract deep acoustic features from MFCC. Then, a LSTM cell is used to capture the characteristics of speech signals from both long-term sequence-dependence relationships and short-term correlations. Followed by the LSTM cell, an attention layer is used to find out the salient emotional regions in a utterance. We refer to the output of ACLN as ALFs.

Meanwhile, we feed the openSMILE features with 384 dimensions into another independent channel of FCN. The designed FCN consists of three fully connected layers and batch normalization layers, where the former is efficient to force correlated human features into low-dimensional representation of speech emotion, while the latter can reduce the internal covariate shift, and speed up the training process [44], [45]. By analogy, we refer to the output of FCN as ELFs.

Unlike the conventional fusion methods [33], where ALFs and ELFs are fused directly and fed into one softmax layer for classification, the extracted ALFs and ELFs by ACLN and FCN are fed into independent softmax layer, followed by specifically designed error feedback propagation of each channel. Until the feature extraction process of each independent channel (ACLN and FCN) converges to a steady state, ALFs and ELFs are then fused in a concatenate layer to complement with each other. With independent feature extraction and error feedback, both ALFs and ELFs, prior to the fusion with each other, can capture the speech emotional state more preciously from the dimensions of their own in comparison with those in [32], [33]. It is thus expected that the fusion of ALFs and ELFs with independent extraction and feedback modification can fully enjoy the complementary advantages of both deep features and empirical features.

### C. INDEPENDENT FEEDBACK MECHANISM

As described above, the design of independent feedback of ACLN and FCN enables a better representation of emotions prior to the fusion of ALFs and ELFs. Basically, the feedback procedure is based on hard example mining and re-training with a specifically designed loss function, as described below.

#### 1) STEP 1: HARD EXAMPLE MINING

We consider hard example mining to classify the samples of a higher classification error rates. The detailed procedures are as follows. First, we consider independent feedback of both ACLN and FCN with CE loss. Until both models converge to steady state, we define the hard-index to distinguish the hard examples within the training database as follows,

$$h_i = \frac{R_i - min(R)}{max(R) - min(R)}, \tag{1}$$

where $R = [R_1, R_2, \cdots, R_N]$ with $N$ denoting the number of training sample, and

$$R_i = -\log_2\left(1 - \frac{m_i}{c_i}\right), \tag{2}$$

is the misclassification ratio information entropy, $m_i$ and $c_i$ in (2) stand for the numbers of misclassification and training epoch of $i$-th sample, respectively. It is noted that the hard index in (2) is normalized by using a max-min function to balance the distributions of samples in training. Clearly, the samples of higher classification error rates can be reasonably viewed as hard examples, i.e., $h_i \geq \beta$, where $\beta \in [0, 1)$ is a non negative number.[1]

The motivation of hard example mining is to enable the re-training of ACLN and FCN by concentrating on the samples of a higher classification error rate (hard examples). In this way, the feature extraction, i.e., ALFs and ELFs, with hard example re-training can better represents the corresponding emotions of hard examples, and thus, can potentially achieve a better performance in comparison with the features without hard example mining.

#### 2) STEP 2: RE-TRAINING WITH HARD EXAMPLE LOSS

In order to emphasize the impact of hard samples, we remove all easy examples from the training database, and re-train the resulting hard examples with a specifically designed loss function, which is referred to as HE loss, given by

$$Loss_i = -h_i(1 - \alpha_i)log(p_i), \tag{3}$$

where $h_i$ and $\alpha_i$ denote the hard-index of sample and ratio of emotion class, respectively, and $p_i$ is the classification error probability. Compared with the conventional loss functions in [41], [46], the proposed HE loss in (3) takes both the distribution of emotion classes and the hard index of samples into account.

To sum up, the independent feedback mechanism with hard example mining and re-training can effectively balance the distribution of utterance label of different emotions, while focusing the contributions of hard examples in the re-training process. It is thus expected that the proposed independent feedback with HE loss yields a superior solution to the existing error back propagation methods [37].

## IV. EXPERIMENTS AND ANALYSIS

### A. EXPERIMENTAL DATABASE

To evaluate the effectiveness of the proposed independent model, we consider three popular databases of both imbalanced and balanced samples, i.e., Interactive Emotional Dyadic Motion Capture (IEMOCAP) [40], Berlin German Emotional Voice Library (EMODB) [47], and Institute of Automation Chinese Academy of Sciences emotion dataset (CASIA) [48].

The IEMOCAP is one of the most widely used databases for SER task that contains extremely imbalanced samples of 10 emotion categories. This corpus is played by 10 skilled speakers in five sessions, and each session is acted by two speakers (1 female and 1 male). In our experiment, we only choose the *improvisations* version, and select 4 emotion categories that totally contain 2280 utterances (angry (289), neutral (1099), happy (284), sad (608)). The duration of all utterances ranges from 3 to 15 seconds sampled at 16kHz.

The EMODB database contains slightly imbalanced samples of 7 emotion categories. The database utterances were produced by 10 professional actors, consisting 5 short-sentences and 5 long-sentences of daily words. All the utterances are saved using the WAV format with a sampling rate and an average time of 16kHz and 2.7 seconds, respectively.

The CASIA contains balanced samples of 6 emotion categories developed by the Institute of Automation, Chinese Academy of Sciences. In our experiment, we choose the text-dependent version of CASIA, where speakers read sentences by using specified emotion. This corpus is acted by 4 professional speakers (2 males and 2 females), and includes 2400 different pronunciations. Basically, all the utterances

[1]The optimal threshold $\beta$ depends crucially on the content of database, and can be obtained directly based on extensive simulations. Specially, we set $h_i = 1$ for the special case of $m_i = c_i$.

**TABLE 1. Parameters of ACLN.**

| Layer | Kernel | Stride | Input | Output |
|---|---|---|---|---|
| Conv block 1 | (5*5)*32 | (1,1) | 1 | 32 |
| Conv block 2 | (5*5)*64 | (1,1) | 32 | 64 |
| Conv block 3 | (5*5)*128 | (1,1) | 64 | 128 |
| Conv block 4 | (5*5)*256 | (1,1) | 128 | 256 |
| Flatten layer | None | None | 256 | 256 |
| LSTM | 128 | None | 256 | 128 |
| Attention layer | 128 | None | 128 | 128 |
| Softmax output | None | None | 128 | 4 |

**TABLE 2. Parameters of FCN.**

| Layer | Weights | Bias | Input | Output |
|---|---|---|---|---|
| Fully connected 1 | (384*256) | 256 | 384 | 256 |
| Batch normalized 1 | None | None | 256 | 256 |
| Fully connected 2 | (256*512) | 512 | 256 | 512 |
| Batch normalized 2 | None | None | 512 | 512 |
| Fully connected 3 | (512*1024) | 1024 | 512 | 1024 |
| Batch normalized 3 | None | None | 1024 | 1024 |
| Softmax output | None | None | 1024 | 4 |

in CASIA are also sampled at 16kHz with a duration of 2 seconds.

### B. EXPERIMENTAL SETUP

The parameters of the independent ACLN and FCN, are listed in Table 1 and Table 2, respectively. As regards the design of ACLN, it contains four convolution blocks, a LSTM cell followed by an attention layer. The FCN consists of three fully connected and batch normalization layers. To avoid over-fitting, a dropout layer with a factor of 0.25 is used before the output layer. We employ SVM as final classifier due to its powerful ability to maximize category interval.[2] The generated discriminative feature representation will be fed into the linear SVM for emotion recognition. In simulations, we adopt random 10-fold cross-validation to conduct the experiments. In EMODB and CASIA, the proportions of training and testing over the database are set to be 80% and 20%, respectively. In IEMOCAP, similar to [19], we conduct speaker-independent experiments, where the sentences from 8 speakers are used for training, and the sentences from remaining 2 speakers are used for testing. We used Adam optimizer in our experiments, and set the initial learning rate as $10^{-3}$. The training epoch and batch size of deep models are set to 1400 and 128. As regards the re-training step of ALFs and ELFs, the stop epochs are set to 300, 50, 80 of IEMOCAP, EMODB and CASIA, respectively, and the re-training epoch set to 800. To be specific, we excluded easy examples at the epoch of 800, and perform re-training with only hard examples and HE loss until the training process converges.

[2]From extensive experiments, which are not included in this study due to space constraint, we validate that SVM can offer the best performance among different classifiers we have tried, namely ELM and MLP. Similar application of SVM for SER is also reported in [49].

### C. COMPARISON: INDEPENDENT TRAINING VS. FUSED TRAINING

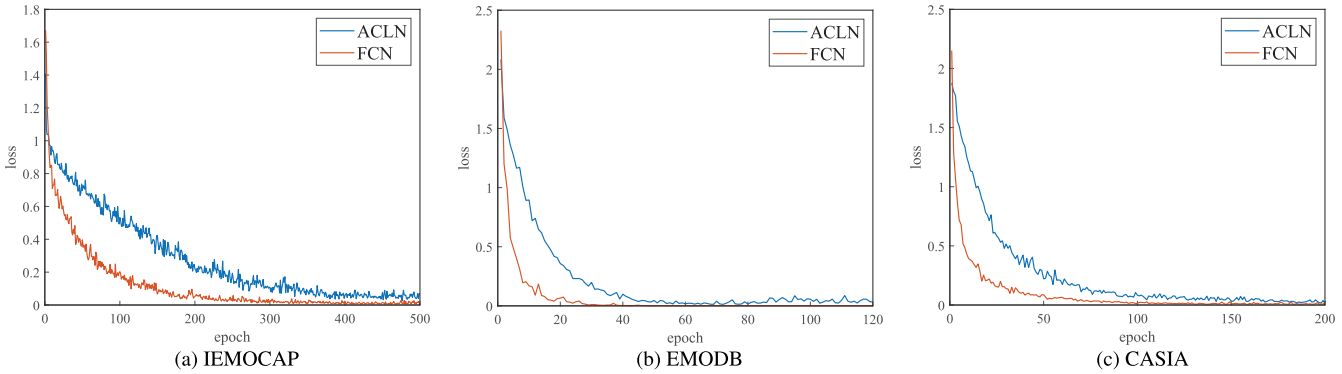
In order to validate the superiority of the proposed independent training framework, we compare the designed independent training with the fused training method [33], where the latter has been considered as a benchmark in related works of feature-fusion. To make the comparisons more convincing, we consider the same experimental setup and list the methods as follows.

- ***Independent training***: The structure of the independent training model is shown in Fig. 1. We feed MFCC and openSMILE features into ACLN and FCN, respectively, enabling the independent training of both ACLN and FCN with different error feedbacks. Until the independent channels converge to steady state, the generated ALFs and ELFs are merged into a vector, and then fed into the SVM for SER.
- ***Fused training***: This is the baseline model of this paper. For fairness of comparison, we feed the same speech signals (MFCC and openSMILE features) into ACLN and FCN to generate ALFs and ELFs, respectively. Unlike the independent training, the ALFs and ELFs are fused through a fully connected layer. With the fused features, the whole network is trained by using the same error feedback propagation [33]. Finally, the fused features are fed into the SVM to distinguish emotions.

We first evaluate the training loss by using ACLN and FCN on three database, i.e., IEMOCAP, EMODB and CASIA. The averaged training loss of both ACLN and FCN in each epoch is shown in Fig. 3. Clearly, we observe that although the loss of ACLN and FCN all converge to the steady state as the increase of iterative epoch training for all databases, the convergence speed of FCN is faster than ACLN for all the emotion databases. More importantly, for tracking an identical task of emotion recognition, the loss of ACLN and FCN of IEMOCAP, EMODB and CASIA in steady state are remarkably different, i.e., the steady-state loss of ACLN and FCN on IEMOCAP are approximately 0.054 and 0.014, respectively. Even for balanced database, i.e., EMODB and CASIA, the steady-state loss of ACLN is approximately 2 times higher than that of FCN. The results shown in Fig. 2 demonstrate that the performance of ALFs and ELFs (extracted by different models, e.g., ACLN and FCN) are different for discriminating emotions. That is, the fusion at the feature-level with one error feedback is unfair, which is neglected by the existing contributions [33].

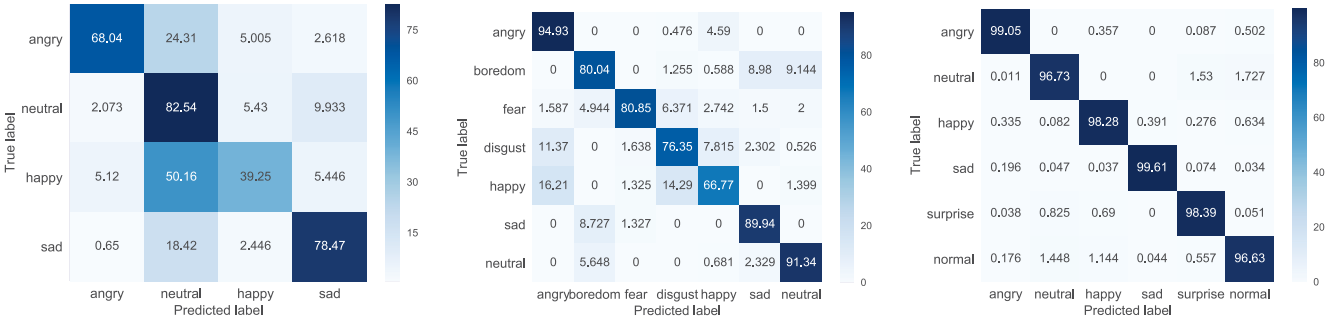
Comparatively, with independent feature extraction, the proposed training mechanism takes the performance difference w.r.t. ALFs and ELFs into consideration, and thus, can fully enjoy the complementary advantages of different features for emotion discrimination.

To gain an insight into the design of independent training, we then conduct an experiment on the comparison between




(a) IEMOCAP (b) EMODB (c) CASIA

**FIGURE 2.** Loss comparison of ACLN and FCN for different databases.




(a) IEMOCAP (b) EMODB (c) CASIA

**FIGURE 3.** Confusion matrix of independent training model without hard example mining and re-training (%).

**TABLE 3.** Performance comparison between independent training and fused training.

| Database | Method | WA(%) | UA(%) |
|---|---|---|---|
| IEMOCAP | Fused training | 69.19 | 61.05 |
| | Independent training | **74.88** | **67.07** |
| EMODB | Fused training | 80.93 | 77.48 |
| | Independent training | **83.33** | **82.81** |
| CASIA | Fused training | 96.14 | 96.14 |
| | Independent training | **98.12** | **98.12** |

the independent training and the fused training. As a rule of thumb, we use weighted accuracy (WA) and unweighted accuracy (UA) as evaluation criteria validate the superiority of the proposed independent training design, as follows.

- ***WA*** - the classification accuracy for the whole test set.
- ***UA*** - the averaged classification accuracy of each emotions.

The evaluation results for IEMOCAP, EMODB and CASIA databases are illustrated in Table 3. From the experiment of imbalanced database, i.e., IEMOCAP, it is seen that the independent training, in comparison with the fused training, achieves notable performance gains in WA and UA of 5.69% and 6.02%, respectively. Even for balanced database, i.e., CASIA, the independent training outperforms the fused training with recognition accuracy gains of approximately 2% in terms of both WA and UA, which validates the advantages of the independent training design. The numerical results are expected since different models (i.e., ACLN and FCN) with independent error feedback make full use of the advantages of extracted features (both ALFs and ELFs), and thus, can generate more discriminative emotion related information in comparison with that of the fused features with the same error feedback.

### *D. HARD EXAMPLE MINING, RE-TRAINING AND LOSS FUNCTION*

To illustrate the effectiveness of the proposed independent error feedback mechanism, we carry out the comparisons between the proposed independent model with/without hard example mining and re-training.

The confusion matrices of classification on IEMOCAP, EMODB and CASIA with and without hard example mining and re-training are shown in Fig. 3 and Fig. 4, respectively. Comparably, it is shown from Fig. 3 that the proposed independent training without hard example mining and re-training achieves an classification accuracy of approximately 98.12% for CASIA database of balanced emotion class. For EMODB database, we acquire the classification accuracy of higher than 80% for most of emotion classes. For


(a) IEMOCAP (b) EMODB (c) CASIA

**FIGURE 4.** Confusion matrix of independent training model with hard example mining and re-training (%).

imbalanced database, i.e., IEMOCAP, however, we observe that the 'happy' emotional state is remarkably confused with the 'neutral' one, and engenders the lowest recognition accuracy of 39.25%. That is, most of the hard examples are concentrated in the 'happy' emotion class. The potential reason lies in the fact of imbalanced distribution across different emotion classes.

For comparison, we evaluate the performance by using the independent training with hard example mining and re-training. As can be observed from Fig. 4, we acquire 2.91% and 4.48% notable improvement in the 'happy' emotion class of IEMOCAP and EMODB, respectively. The results are expected since the hard example re-training could extract more valuable features from the samples of higher classification error rate so that we can identify them better. In addition, we gain a slight improvement of 1.98% in terms of recognition accuracy for the balanced database of CASIA.

From the result of the above contrast experiment, we observe that the proposed hard example mining and re-training is capable of improving the recognition accuracy for imbalanced database, i.e., IEMOCAP. Therefore, we conduct an experiment on IEMOCAP to evaluate the impact that the hard example mining and re-training has on the distribution of emotion classes. Fig. 5 plots the numbers of hard examples and easy examples in IEMOCAP. As can be observed from Fig. 5, the ratios of easy example and hard example are 49.4% and 50.6%, respectively. Without hard example mining, the 'neutral' emotion class in IEMOCAP contains the largest number of samples, which is almost 4 times higher than that of 'happy' emotion class. Using the proposed hard example mining, the ratio of samples between the above two emotion class is reduced to 2.17, demonstrating that the distributions of emotion classes is balanced. We further list the ratios of emotion class distribution with and without hard example mining and re-training in Table 4. As shown in Table 4, we can see that by removing all easy examples using hard example mining, the proportion of 'happy' emotion class (the lowest ratio of IEMOCAP emotion classes) is remarkably increased by 7.28%, i.e., from 12.55% to 19.83%, while the proportion of 'neutral' emotion class reduces from 47.53% to 43.01%. We plot the

**TABLE 4.** Emotion class distribution of IEMOCAP - with/without hard example mining and re-training.

| | Angry(%) | Neutral(%) | Happy(%) | Sad(%) |
|---|---|---|---|---|
| W/O hard example re-training | 13.05 | 47.53 | 12.55 | 26.87 |
| W/ hard example re-training | 14.08 | 43.01 | 19.83 | 23.08 |

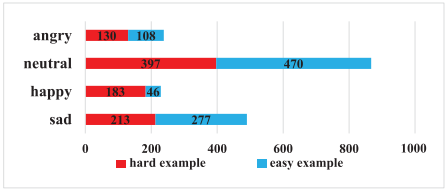



**FIGURE 5.** Numbers of hard example and easy example in IEMOCAP.

SER accuracy with and without hard example re-training in IEMOCAP. It is seen from Fig. 6 that the SER performance by using hard example re-training in terms of both UA and WA in IEMOCAP is more robust to that without re-training. The results are consistent with that shown in Fig. 5 and Table 4, demonstrating that the proposed HE-aided re-training is able to balance the emotion class for unbalanced database, and thus, has the potential of improving the recognition performance in practice. More importantly, the re-training process is based on the resulting hard examples of all emotion classes. In this way, the re-training with the generated ALFs and ELFs can effectively characterize the expressions of hard examples, and thereby, has the potential of achieving a higher classification accuracy in comparison with the conventional training scheme [41].

Next, we evaluate the performance of the proposed HE loss, and consider both CE loss and focal loss as the baselines in simulation. For fairness of comparison, three loss functions are applied in the proposed independent training model for IEMOCAP, EMODB and CASIA databases. As shown in Table 5, the proposed HE loss, in comparison with CE loss and focal loss, yields a superior solution to all databases

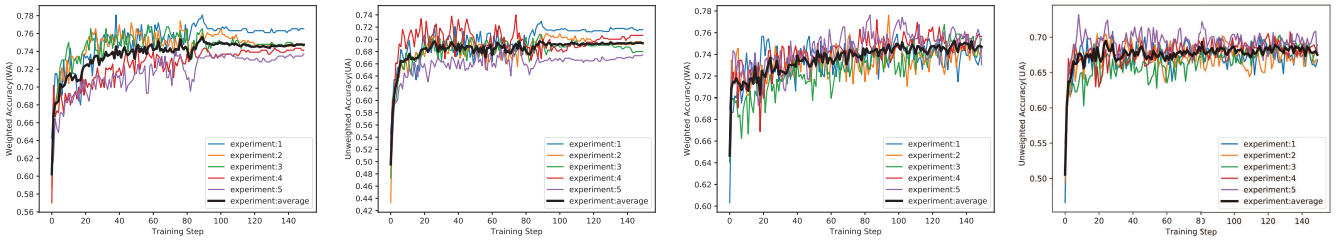



**FIGURE 6.** Recognition curves of IEMOCAP across 5 experimental runs. (a) WA with hard example re-training. (b) UA with hard example re-training. (c) WA without hard example re-training. (d) UA without hard example re-training.

**TABLE 5.** Comparison of different loss functions.

| Database | Loss | WA(%) | UA(%) |
|---|---|---|---|
| IEMOCAP | CE loss | 74.88 | 67.07 |
| | focal loss | 74.72 | 67.74 |
| | HE loss | **74.98** | **68.83** |
| EMODB | CE loss | 83.33 | 82.88 |
| | focal loss | 83.51 | 83.56 |
| | HE loss | **85.76** | **86.12** |
| CASIA | CE loss | 98.12 | 98.12 |
| | focal loss | 98.08 | 98.08 |
| | HE loss | **98.15** | **98.15** |

in terms of both WA and UA. Since UA has take the data distribution into account, the significance is even more pronounced by improving UA in SER with data imbalance [30]. It is seen from Table 5 that the proposed HE loss achieves absolute improvement of 1.09% (from 67.74% to 68.83%) and 1.76% (from 67.07% to 68.83%) compared with focal loss and CE loss, in terms of UA for IEMOCAP, and remarkably outperforms both CE loss and focal loss for EMODB by an absolute improvement of 2.56% and 3.24%, respectively. Even for balanced database CASIA, the proposed HE loss still performs better than the conventional baselines in terms of both WA and UA. The promising result of the proposed method reflects the effectiveness of the hard example mining and re-training mechanism with HE loss.

To further evaluate the performance of the hard-example loss w.r.t. the emotion recognition for imbalanced database, we list the comparison of loss functions in terms of recognition accuracy of each emotion class for IEMOCAP database. As shown in Table 6, we can see that the 'neutral' and 'happy' emotion classes have the highest and the lowest recognition accuracy, respectively, for all loss functions. The result is expected since the proportions of such two emotion classes are of the largest and the smallest ones in IEMOCAP database, respectively. By using the proposed HE loss as the feedback of training, we observe that the recognition accuracy of 'angry', 'happy' and 'sad' emotion classes are increased by 0.19%, 1.02% and 3.29% in comparison with that using focal loss. The result is consistent with that of Table 4, where the emotion classes are balanced by removing easy examples during the re-training process. Although the

**TABLE 6.** Detail performance of IEMOCAP training under different loss functions.

| Loss | Angry(%) | Neutral(%) | Happy(%) | Sad(%) |
|---|---|---|---|---|
| CE loss | 68.04 | **82.54** | 39.25 | 78.46 |
| Focal loss | 70.36 | 80.31 | 41.14 | 79.17 |
| HE loss | **70.55** | 80.18 | **42.16** | **82.46** |

recognition accuracy of utterances with labels of 'neutral' is slightly reduced by 0.13%, the performance of the proposed HE loss, as demonstrated by Table 5, is still remarkably better than those of the baselines, i.e., CE loss and focal loss functions.

### *E. COMPARISON WITH THE STATE-OF-THE-ART METHODS*

To verify the superiority of the proposed independent feedback mechanism, we provide the contrast experiment between the proposed independent training framework (with independent feedback mechanism) and several state-of-the-art contributions for IEMOCAP, EMODB and CASIA databases. As shown in Table 7, for IEMOCAP database, the proposed independent training framework outperforms the conventional feature fusion methods [32], [33], demonstrating the superiority of the independent design of feature extraction and feedback mechanism. Furthermore, in comparison with the latest contribution that aims at addressing the problem of data imbalance [37], our method can provide a higher classification accuracy in terms of both WA and UA. The results are expected. On one hand, the feature extractions of ALFs and ELFs are independent, and thus, enable ALFs and ELFs to characterize effective emotions from different perspective without suffering from the penalty of loss difference in the conventional feature-fusion works. On the other hand, taking hard examples into consideration, the proposed independent feedback mechanism not only enables a sufficient training on hard examples, but also is beneficial to balance emotion classes.

For EMODB database, although the proposed independent training framework performs slightly inferior to that of [33], our work achieves a superior performance to those of other methods in terms of both WA and UA in IEMOCA and

**TABLE 7. Comparison with state-of-the-art methods.**

| Database | Ref. | WA(%) | UA(%) |
|---|---|---|---|
| IEMOCAP | [33] | 57.99 | 56.55 |
| | [32] | 60.35 | 63.98 |
| | [37] | 74.60 | 66.70 |
| | Proposed | **74.98** | **68.83** |
| EMODB | [50] | 82.32 | 80.51 |
| | [27] | 82.82 | 82.14 |
| | [33] | **87.85** | **87.49** |
| | Proposed | 85.76 | 86.12 |
| CASIA | [51] | 94.60 | 94.60 |
| | [52] | 95.80 | 95.80 |
| | Proposed | **98.15** | **98.15** |

CASIA.[3] For example, the proposed independent framework outperforms the state-of-the-art model of 3D CRNN with attention [27], by an absolute improvement of 2.94% (from 82.82% to 85.76%) and 3.98% (from 82.14% to 86.12%) in terms of both WA and UA. In addition, as a comparison with balanced database, i.e., CASIA, our method achieves an obvious recognition accuracy gain of 2.35% in terms of WA and UA, respectively. The above results demonstrate that the proposed independent training framework outperforms the conventional state-of-the-art solutions for both balanced and imbalanced databases.

## V. CONCLUSION AND PERSPECTIVE

This paper proposed an independent training framework for SER. Specifically, in order to take full advantages of both deep features and empirical features, we designed an independent feature extraction for both ALFs and ELFs by feeding MFCC and openSMILE features into ACLN and FCN. In addition, we proposed hard example mining-based re-training mechanism with HE loss in the feedback process, which focus learning of each independent model mainly on hard negative examples. With the independently extracted features of both ALFs and ELFs, a SVM classifier is adopted to distinguish emotions. Experimental results on three public databases show that the proposed independent training outperforms the state-of-the-art feature-fusion methods. The results also demonstrated that the independent feature extraction of ALFs and ELF can better represent the emotions from different perspective in comparison with that of the fused training. Furthermore the independent feedback using hard example mining-based re-training mechanism remarkably improves the feature learning ability of each independent model, which can generate more discriminating emotion related feature representations than the conventional error feedback methods.

[3]The potential reason of the performance inferiority to [33] is due to the different inputs of deep model, i.e. the input employed by [33] is spectrogram while the proposed scheme used MFCC from the aspect of computational efficiency.

## REFERENCES

[1] S. Pathak and V. Kolhe, "A survey on emotion recognition from speech signal," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 5, no. 7, pp. 447–450, 2016.
[2] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, Mar. 2018.
[3] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommun. Syst.*, vol. 52, no. 3, pp. 1467–1478, Mar. 2013.
[4] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.
[5] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
[6] T.-W. Sun, "End-to-End speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020.
[7] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020.
[8] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1675–1685, Nov. 2019.
[9] K. Wang, N. An, B. Nan Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, Jan. 2015.
[10] 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.
[11] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, Feb. 2015.
[12] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. 17–20.
[13] J.-C. Wang, Y.-H. Chin, B.-W. Chen, C.-H. Lin, and C.-H. Wu, "Speech emotion verification using emotion variance modeling and discriminant scale-frequency maps," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1552–1562, Oct. 2015.
[14] S. Kuchibhotla, H. D. Vankayalapati, and K. R. Anne, "An optimal two stage feature selection for speech emotion recognition using acoustic features," *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 657–667, Dec. 2016.
[15] B. Schuller, S. Steidl, A. Batliner, and A. E. A. Vinciarelli, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, Aug. 2013, pp. 148–152.
[16] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, 2013, pp. 3–10.
[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
[18] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[19] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[20] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.
[21] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[22] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5089–5093.

[23] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, Mar. 2019.
[24] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks," *Int. J. Speech Technol.*, vol. 22, no. 3, pp. 497–510, Sep. 2019.
[25] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.
[26] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Process.*, vol. 12, no. 6, pp. 713–721, Aug. 2018.
[27] 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.
[28] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16560–16572, 2020.
[29] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019.
[30] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.
[31] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2666–2670.
[32] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. Interspeech*, Sep. 2018, pp. 152–156.
[33] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75798–75809, 2019.
[34] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.
[35] P.-Y. Shih, C.-P. Chen, and H.-M. Wang, "Speech emotion recognition with skew-robust neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2751–2755.
[36] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using GANs for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 171–175.
[37] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Focal loss based residual convolutional neural network for speech emotion recognition," 2019, *arXiv:1906.05682*. [Online]. Available: http://arxiv.org/abs/1906.05682
[38] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5150–5154.
[39] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech*, Aug. 2017, pp. 1089–1093.
[40] 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, pp. 335–359, Dec. 2008.
[41] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[42] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 1459–1462.
[43] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc INTERSPEECH*, Jan. 2009, pp. 312–315.
[44] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6675–6679.
[45] S. Wu, G. Li, L. Deng, L. Liu, D. Wu, Y. Xie, and L. Shi, "$L1$-Norm batch normalization for efficient training of deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2043–2051, Jul. 2019.
[46] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.
[47] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Sep. 2005, pp. 1517–1520.
[48] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: A chinese natural emotional audio–visual database," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 6, pp. 913–924, Nov. 2017.
[49] S. Zhang, A. Chen, W. Guo, Y. Cui, X. Zhao, and L. Liu, "Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition," *IEEE Access*, vol. 8, pp. 23496–23505, 2020.
[50] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–9, Mar. 2017.
[51] W. Zhang, D. Zhao, Z. Chai, L. T. Yang, X. Liu, F. Gong, and S. Yang, "Deep learning and SVM-based emotion recognition from chinese speech for smart affective services," *Softw. Pract. Exper.*, vol. 47, no. 8, pp. 1127–1138, 2017.
[52] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN," *Sensors*, vol. 17, no. 7, pp. 1694–1707, 2017.

**SHUNMING ZHONG** received the B.S. degree in electronic information science and technology from the Shaoguan College, Guangdong, China, in 2018. He is currently pursuing the M.E. degree in electromagnetic field and microwave technology with the School of Physics and Telecommunications Engineering, South China Normal University, Guangzhou, China. His research interests include spoken signal processing, deep neural networks, and speech emotion recognition.

**BAOXIAN YU** received the B.E. degree in communication engineering and the Ph.D. degree in information and communication engineering from Sun Yat-sen University, Guangzhou, China, in 2014 and 2019, respectively. From 2017 to 2018, he was a Research Assistant with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. Since 2019, he has been with the Department of Physics and Telecommunications Engineering, South China Normal University, Guangzhou, China, where he is currently a Research Associate. His research interests include signal processing for high spectral efficiency communications, high speed optical transmissions, and biomedical engineering.

**HAN ZHANG** (Member, IEEE) received the Ph.D. degree from the School of Information Science Technology, Sun Yat-sen University, China. Since 2009, he has been with the Department of Physics and Telecommunications Engineering, South China Normal University, China, where he is currently a Professor. From 2012 to 2013, he was a Senior Research Associate with the Department of Electrical Engineering, City University of Hong Kong. He holds the positions of the Chief Scientist and a Technical Advisor in several high-technology enterprises in China. His research interests include signal processing for the Internet of Things and biomedical engineering.