Expert Systems With Applications 214 (2023) 118943

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Learning multi-scale features for speech emotion recognition with connection attention mechanism

Zengzhao Chen [a,c,*], Jiawen Li [a,b], Hai Liu [a,c], Xuyang Wang [d], Hu Wang [a,c], Qiuyu Zheng [a,b]

[a] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
[b] National Engineering Laboratory For Educational Big Data, Central China Normal University, Wuhan 430079, China
[c] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
[d] Luoyang Institute of Electro-Optical Equipment, Aviation Industry Corporation, Luoyang, 471023, Henan, China

ARTICLE INFO

Keywords:
Speech emotion recognition
Features fusion
Connection attention mechanism
Frame-level features
Utterance-level features

ABSTRACT

Speech emotion recognition (SER) has become a crucial topic in the field of human–computer interactions. Feature representation plays an important role in SER, but there are still many challenges in feature representation such as the inability to predict which features are most effective for SER and the cultural differences in emotion expression. Most previous studies use a single type of feature for the recognition task or conduct early fusion of features. However, a single type of feature cannot well reflect the emotions of speech signals. Also, different features contain different information, direct fusion cannot integrate the advantages of different features. To overcome these challenges, this paper proposes a parallel network for multi-scale SER based on a connection attention mechanism (AMSNet). AMSNet fuses fine-grained frame-level manual features with coarse-grained utterance-level deep features. Meanwhile, it adopts different speech emotion feature extraction modules according to the temporal and spatial features of speech signals, which enriches features and improves feature characterization. The network consists of a frame-level representation learning module (FRLM) based on the time structure and an utterance-level representation learning module (URLM) based on the global structure. Besides, improved attention-based long short-term memory (LSTM) is introduced into FRLM to focus on the frames that contribute more to the final emotion recognition result. In URLM, a convolutional neural network with the squeeze-and-excitation block (SCNN) is introduced to extract deep features. In addition, the connection attention mechanism is proposed for feature fusion, which applies different weights to different features. Extensive experiments are conducted on the IEMOCAP and EmoDB datasets, and the results demonstrate the effectiveness and performance superiority of AMSNet. Our code will be publicly available at https://codeocean.com/capsule/8636967/tree/v1.

1. Introduction

Human emotional state is a significant factor in human communication, and speech plays an important role in emotional expression (Dolan, 2002). Speech emotion recognition (SER) aims to recognize and classify emotional states by extracting and modeling speech features, such as short-term energy, zero-crossing rate, pitch, formant, duration, prosody, spectral features, Mel frequency cepstral coefficient (MFCC) features (Jiang & Cai, 2004; Sahu, 2019). The difference of emotions expressed in speech is related to the distribution of the time structure, amplitude structure, fundamental frequency structure, and formant structure. The acoustic features of speech, including prosodic features, spectral features and other features, can more intuitively reflect the differences between these features. In recent years, SER is applied to many fields, such as human–computer interaction (Cowie et al., 2001), emotion learning detection, mental health analysis, and customer service detection.

The first problem that needs to be investigated for SER is which features are effective for emotion recognition. Although quite a few features have been identified as highly associated with emotion recognition, this problem is not studied in-depth. The second problem is how to identify emotions based on the extracted features.

Traditional SER methods mainly use machine learning such as support vector machine (Lin & Wei, 2005), naive Bayesian model (Wang, An, Li, Zhang, & Li, 2015), and k-nearest neighbor (Lanjewar, Mathurkar, & Patel, 2015) with hand-crafted features. Meanwhile, various emotion feature sets such as INTERSPEECH 2013 (Schuller et al.,

[*] Corresponding author at: Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China.
E-mail addresses: zzchen@ccnu.edu.cn (Z. Chen), jiawenli@mails.ccnu.edu.cn (J. Li), hailiu0204@ccnu.edu.cn (H. Liu), wxysunshy@163.com (X. Wang), wanghu@mails.ccnu.edu.cn (H. Wang), qiuyu@mails.ccnu.edu.cn (Q. Zheng).

https://doi.org/10.1016/j.eswa.2022.118943
Received 30 March 2022; Received in revised form 14 September 2022; Accepted 28 September 2022
Available online 8 October 2022

2013), A VEC-2016 (Valstar et al., 2016), and GeMAPs (Eyben et al., 2015) have been used. However, the performance of such hand-crafted features is limited, and the increase of the feature set size also increases the training complexity.

Recently, with the rapid development of deep learning (Liu et al., 2022, 2022), some techniques have been applied to SER and achieved significant performance improvements (Fayek, Lech, & Cavedon, 2017; Kwon et al., 2021; Xie et al., 2019). Also, numerous researchers have made efforts in the research of the above problems. The application of neural networks makes it possible to solve the problem of feature selection. The method proposed by Han, Yu, and Tashev (2014) is representative. It uses a deep neural network model to learn deep features from the input features of pitch period, harmonic-to-noise ratio, and MFCCs. Satt, Rozenberg, and Hoory (2017) presents a novel framework based on the convolutional neural network (CNN) that uses Mel-scale spectrograms as input. Compared with hand-crafted features, a spectrogram is a primitive representation without much specific feature representation. Neural networks can automatically extract features from the original representation, thus reducing the overhead of feature engineering. The subsequent work (Hou, Li, & Lu, 2020; Meng, Yan, Yuan, & Wei, 2019; Wu et al., 2021) directly applied the convolution layer and pooling layer to the spectrograms and achieved good results, which proves the advantage of using spectrograms as input features.

The CNN model or long short-term memory (LSTM) model is commonly used in SER tasks based on automatically learned features. A two-stage CNN was proposed by Mao, Dong, Huang, and Zhan (2014) to obtain more useful representations of emotional features, but it still cannot solve the problem of dynamic features of speech signals. Chen, He, Yang, and Zhang (2018) reduced the influence of non-verbal factors on the recognition results by calculating deltas and delta-deltas for the log-Mels and proposed a 3-D attention-based convolutional recurrent neural network to better capture the time–frequency relationship of the features. To learn more emotional details of different features, a parallel network model was proposed by Jiang, Fu, Tao, Lei, and Zhao (2019), in which CNN was used for Log-Mel spectrograms and LSTM was used for frame-level features. It is clear that deep features and hand-craft features have their own advantages in SER tasks. Although automatically learning features by deep models can capture significant emotional information of speech signals, the above work fails to show the performance improvement of fusing hand-craft features and automatically learned features (see Fig. 1).

In recent years, some researchers have focused on the fusion of hand-craft features and automatically learned features to take advantage of both features. The fusion method was introduced by Guo, Wang, Dang, Liu, and Guan (2019) and Kumaran, Radha Rammohan, Nagarajan, and Prathik (2021). The method using two types of features improves SER performance compared to that using one type of feature. However, fusion of different type of features such as prosodic features and spectral features has been ignored by some existing studies. The richness of features has extremely positive impact on the experimental results has been proved in Guo, Wang, Dang, Zhang, and Guan (2018) and Luo, Zou, and Huang (2018). Acoustic features and prosodic features describe speech signals from different aspects and are complementary. The fusion of two features can obtain richer emotional information from speech. Prosody features are extracted in time domain. The discrete Fourier transform is used to transform the time domain into the frequency domain, and the frequency domain representation of the speech signal is generated, so the deep neural network is used to process the speech spectrum for high-dimensional feature extraction. The fusion of the two can enrich the characterization of features. It is further explained that models with two different features have different performances for the same task. A lot of previous work has connected different networks serially for emotion recognition. It is easier to design models suitable for segment-level features. However, due to the inheritance relationship between models, sequential model structures may lose some emotional information. The final problem is designing the appropriate inputs to meet the needs of the different models. To maximize the contributions of hand-craft features and deep features in SER tasks, different models are used in this paper, and an attention mechanism is adopted for model fusion. Specifically, training the two types of features separately can generate different highly abstract feature representations, which can better integrate the advantages of the two types of features. Since the contributions of hand-craft features and deep features to the final recognition results are different, the use of an attention mechanism can highlight the importance of features. The weight value is used to adjust the importance degree of the two features, and the relatively more important features are given a higher weight so that the features have a larger contribution to the overall recognition effect, thereby significantly improving the model performance.

The major contributions of this work are summarized as follows:

i. An independent training method is proposed to overcome the limitation of traditional feature fusion at the feature level. A parallel connection mode is conducted with multiple features as input, the complete emotional details of different functional features are learned simultaneously. While Deep features can better express the relationship between frequency and time while hand-craft features can comprehensively describe the basic information of speech signals. Thus, different network structures can be used for feature extraction to combine the advantages of spatial and temporal features of signals.
ii. The connection attention mechanism for network fusion is proposed, which assigns different weight values to different types of features. By using the weight-based attention mechanism, the advantages of various features can be integrated, and the emotion recognition ability of the model can be improved.
iii. The proposed method is evaluated on two open datasets, i.e., EmoDB, and Interactive Emotional Dyadic Motion Capture Database (IEMOCAP). It obtains a weighted accuracy of 69.22% and 88.34% on the two datasets, showing the effectiveness of the proposed method.

The rest of this paper is organized as follows. A brief review of relevant work is given in Section 2. Section 3 describes the proposed method of fusing hand-craft features and deep features based on the attention mechanism. The datasets, feature extraction, details of experiments, and analysis of the experimental results are presented in Section 4. Section 5 concludes this paper.

2. Related work

2.1. Speech emotion recognition

SER mainly includes three modules: speech signal preprocessing, feature extraction, and emotion classification. The preprocessing stage is a prerequisite for the following feature extraction. After processing, the original speech signal can be transformed into a digital quantity that is more conducive to special transformation extraction, which improves the robustness of SER. For example, the instantaneous energy of the signal is calculated by Malik, Malik, Mehmood, and Makhdoom (2021), which has a better processing effect for noisy speech signals. Feature extraction is a necessary step in identifying discriminant features corresponding to different types of emotions. Traditional deep learning network structures such as CNN (Zhao, Mao, & Chen, 2018) and RNN (Kumaran et al., 2021) have been widely used in SER. With the development of deep learning, some popular networks such as sequential capsule networks (Wu et al., 2021) are also used for SER to extract more spatial information and further improve the ability of the model to capture contextual information. However, these networks cannot capture context information over long distances. In Chen and Huang (2021), a dual network based on LSTM is introduced to solve

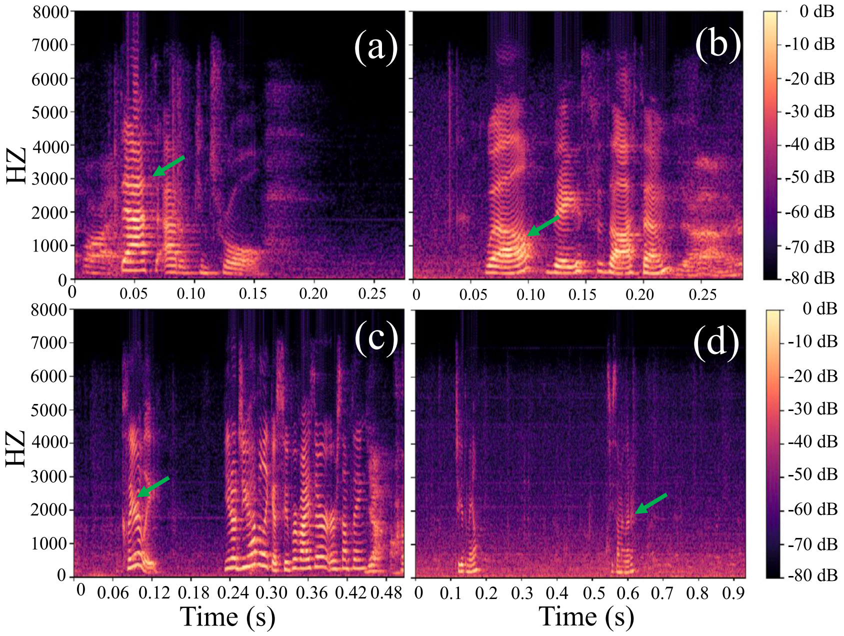


Fig. 1. Spectrograms of different emotions. From top to bottom and from left to right are the spectrograms of angry, happy, neutral and sad. The vertical axis of the spectrogram corresponds to frequency, the horizontal axis corresponds to time, and the color of the image corresponds to energy. The phonetic and energy differences of different emotions can be reflected in the color of the spectrogram.

Table 1
Interspeech 2020 paralinguistics challenge feature set.

Low-level descriptors	Functionals
PCM loudness	Position maximum/minimum
MFCC [0–14]	Arithmetic mean, Standard deviation
Log Mel Freq. Band [0–7]	Skewness, Kurtosis
LSP Frequency [0–7]	Linear regression coefficients 1/2
F0 by sub-Harmonic sum	Linear regression error Q/A
F0 envelop	Quartile 1/2/3
Voicing probability	Quartile range 2-1/3–2/3-1
Jitter local	Percentile 1.0/99.0
Jitter DDP	PCTLrange[a] 0-1/99-1
Shimmer local	Up-level time 75/90

[a]PCTLrange [] (string) [default: 0–1] means that “Array that specifies which inter percentile ranges to compute. A range is specified as ‘n1–n2’ (where n1 and n2 are the indices of the percentiles as they appear in the percentile[] array, starting at 0 with the index of the first percentile).”

the problem. At the same time, unsupervised representation learning catches the attention of researchers. The Transformer-based (Zhang et al., 2021) encoders is used which can be pre-trained with large amounts of unlabeled audio from a variety of datasets and enables learning of more general and robust acoustic representations. Contrastive predictive coding (Li et al., 2021) is used to learn significant representations in unlabeled datasets. Adding noisy data is proved a feasible solution (Xu, Zhang, & Zhang, 2021) when high quality data is insufficient. It also improves the robustness of the model. For emotion classification, in addition to traditional machine learning methods, end-to-end learning methods (Kumar et al., 2021) are now also applied to SER.

2.2. Attention mechanism

The function of the attention mechanism in image processing and machine translation is to enable the model focus on the key information that contributes more to the recognition results during training (Bahdanau, Cho, & Bengio, 2014). Considering that not all speech frames contain emotional information in SER, making the model learn useful keyframes can greatly improve the training efficiency and recognition results. In SER, the combination of attention mechanism with CNN (Zhang, Du, Wang, Zhang, & Tu, 2018), 3-D CNN (Zhao et al., 2021), LSTM (Li, Liu, Yang, Sun and Wang, 2021), and other models has achieved good results.

2.3. Acoustic features

Since speech signal is a short-term stable signal, it is processed into frames, and each frame is treated as a stationary signal. Meanwhile, to make the change between frames natural, the adjacent frames need to overlap with each other.

Open-source Speech and Music Interpretation by Large-space Extraction (openSMILE) was developed by Schuller et al. (2011) from the Technical University of Munich for audio feature extraction and speech signal classification. It has been widely used in automatic emotion recognition tasks.

As shown in Table 1, the Interspeech 2020 Paralinguistics Challenge Feature Set contains 1582 features from the base of 34 low-level descriptors (LLDs). There are 34 corresponding delta coefficients attached, and 21 generic functions are applied to the profiles of 68 LLDs. In addition, 19 statistical functions are applied to the four tonal LLDs and their four first-order differences, and the set of the above 21 statistical functions excluding percentile1.0 and PCTLrange0-1.

3. Proposed AMSNet method

3.1. Overview of AMSNet

In this section, the details of AMSNet is introduced in two parts. To automatically learn the optimal spatio-temporal representation of speech signals and provide granular emotion analysis, this study constructs two modules to extract different levels of emotion representation, i.e., FRLM and URLM. CNN and LSTM models show superiority in capturing detailed spatial and temporal information, respectively. Therefore, improved CNN and LSTM models are introduced in URLM and FRLM.

FRLM is composed of Bi-directional long short-term memory network (BLSTM), and the extracted features are LLDs. It mainly focuses on

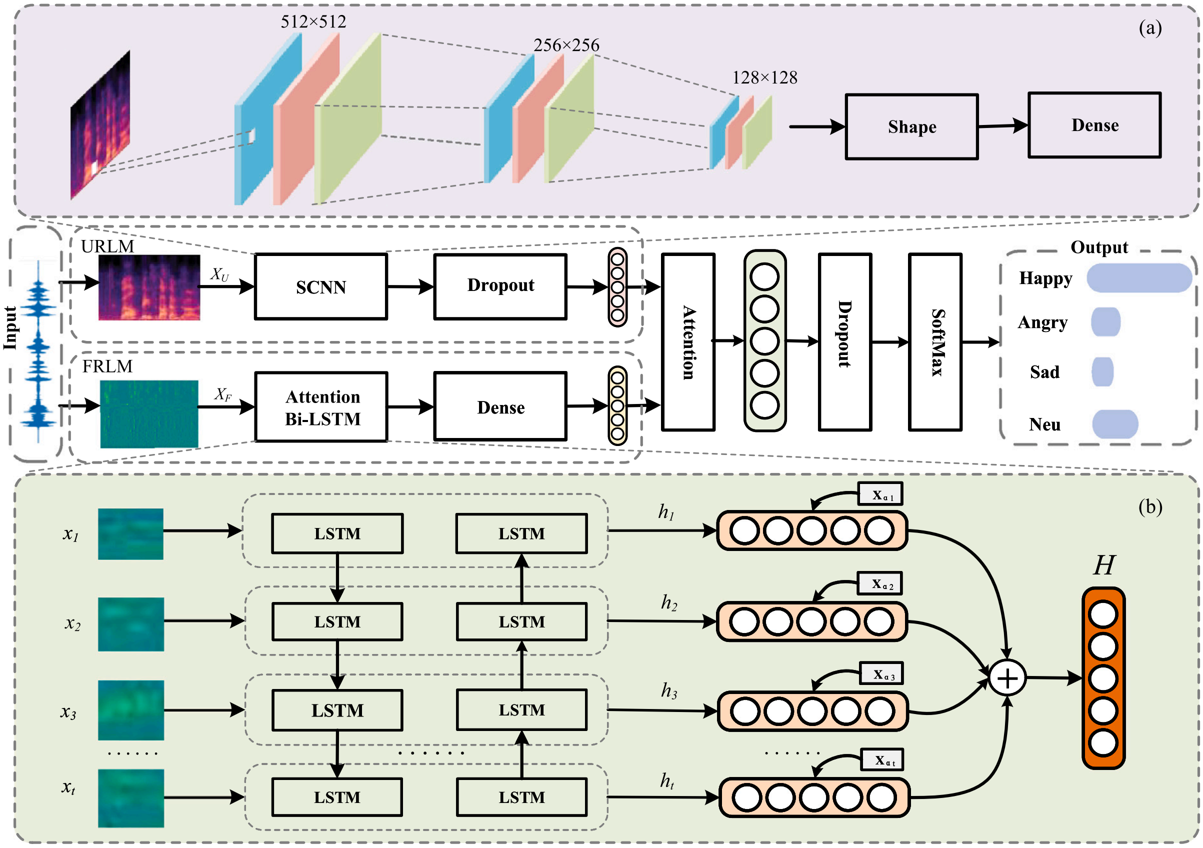


Fig. 2. Structure of proposed AMSNet. (a) Details of SCRNN in URLM. (b) Details of Attention BLSTM in FRLM.

learning global context information from frame-level temporal features and analyzing the fine-grained composition of emotions from feature changes between frames. URLM is composed of spatial CNN, which is mainly used for extracting features from spectrograms and removing the interference from the deep features extracted from utterance, thus learning the coarse-grained features of speech emotion efficiently. Finally, an attention mechanism is used to fuse the parallel networks and assign different weights to both two kinds of features.

Based on this, AMSNet is constructed to extract features from different levels of speech signals and fuse them for multi-scale feature representation, thus retaining the advantages of different types of features.

3.2. Frame-level representation learning module

In this section, a network module based on LSTM sequence structure is proposed to better learn the high-level feature representation with temporal correlation characteristics from frame-level features. And the structure will be introduced in detail. Context information is also important for SER tasks, and the standard LSTM model cannot capture such information, so BLSTM is introduced in FRLM.

Meanwhile, a variant network introduced by Graves, Jaitly, and Mohamed (2013) is adopted, which adds weighted peephole connections from the constant error carousel to the gates of the same memory block. By generating gate degrees directly with the current cell state, the peephole connections allow all gates to inspect into the cell even when the output gate is closed. The improved LSTM model also consists of four parts. The forgetting gate that determines whether to discard information by judging the importance of the current input information denotes g^f. The input of the current unit denotes x_t, and it is the output of the previous unit h_{t-1}. After passing through the sigmoid layer, the output g^f falling within 0 and 1 is obtained. The input gate g^i determines whether the unit state needs to be updated by judging the last output and the current input information. The output gate g^o determines the extent to which the current output depends on the current memory cell state.

The BLSTM model is improved on the unidirectional LSTM network by introducing a second layer, which combines forward LSTM and backward LSTM. Therefore, the model can store the information from past and future. As shown in Fig. 2(b), the model has two sub-networks of context, and the kth output indicates

$$h_k = [\overrightarrow{h_k} \oplus \overleftarrow{h_k}], \tag{1}$$

where $\overrightarrow{h_k}$ represents the output to the forward LSTM layer, and $\overleftarrow{h_k}$ represents the output to the backward LSTM layer.

The input of the BLSTM model is the utterance-level feature of each signal $Z = \{z_1, z_2, \ldots, z_T\}(Z \in R^{d^u \times T})$, where d^u means the number of utterances in the speech signal, and T indicates the sentence length. Each $Z_i(i = 1, 2, \ldots, T)$ denotes fed into a BLSTM to encode the features, and the final output is $H = \{h_1, h_2, \ldots, h_T\}$

Since all the input feature vectors have different effects on emotion recognition, the attention mechanism is introduced to assign weights to different features.

Let $H \in R^{d \times T}$ be a matrix composed of the output vector $H = \{h_1, h_2, \ldots, h_T\}$ generated by the LSTM layer. The representation r of the sentence consists of the weighted sum of these output vectors,

$$S = \tanh(H), \tag{2}$$

$$\alpha = softmax(w^T S), \tag{3}$$

$$r = H\alpha^T, \tag{4}$$

where w^T indicates a transpose of the trained parameter vector. The dimension of w, α, and r denotes respectively d^u, T, and d^u.

To prevent overfitting, the dropout layer is placed after the attention layer. Finally, the output goes through the full connection layer.

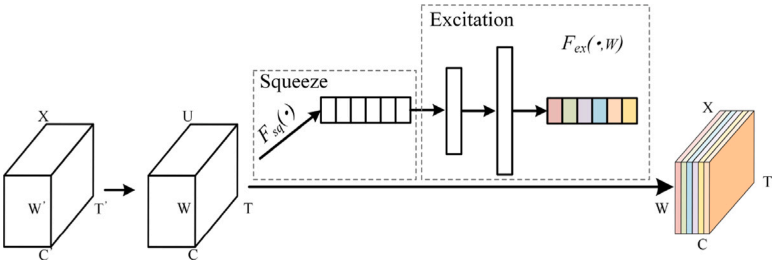


Fig. 3. Computation of the temporal SE-block.

3.3. Utterance-level representation learning module

Due to the large dimension and rich information of utterance-level features, the squeeze-and-excitation block (SE-block) is introduced to apply an attention mechanism to URLM so that more important feature information obtains a higher weight. After the speech signal to recognize is input, the model starts to extract utterance-level features, which are then fed into the feature representation module.

URLM is composed of SCNN. The SCNN block consists of three stacked time convolution blocks (Wang et al., 2021), and each convolution block contains a convolution layer. The first two convolution blocks are ended with SE-block. Fig. 3 shows the computation involved in the SE-block.

The SE-block is an addition to the CNN block, and it adaptively recalibrates the input feature maps. The input and output of the convolution operation are $X(W' \times T' \times C')$ and $R(W \times T \times C)$, respectively. The SE-block can map input $X \in \mathbb{R}^{W' \times T' \times C'}$ to transformations of features $R \in \mathbb{R}^{W \times T \times C}$. In the following calculation, the output is denoted as $R = [r_1, r_2, \dots, r_C]$, while $r_k (k \in [1, 2, \dots, C])$ is calculated by

$$r_k = V_k * X = \sum_{s=1}^{C'} V_k^s * X^s, \tag{5}$$

where $*$ denotes the convolution operation, $V = [v_1, v_2, \dots, v_C]$ is the parameter of the kth convolution kernel; $X = [x_1, x_2, \dots, x_C]$ and $r_k \in \mathbb{R}^{W \times T}$ represent a two-dimensional space kernel, representing a single channel V_k acting on the corresponding channel X. The output is generated by summing up all channels. After inputting the spatial features of a channel, the feature space relations are learned. However, since the convolution results of each channel are directly added, the channel feature relations are mixed with the spatial relations learned by the convolution kernel. The traditional convolution operation only fuses the features of a local region space and does not fuse the features between channels. The SE-block uses backpropagation to learn the weight coefficient of each feature channel, which represents the importance of the channel. According to the weight coefficient, the feature information of the channel is extracted to achieve feature fusion between channels and improve network performance.

The squeeze operation $F_{sq}(r_k)$ compresses a two-dimensional channel into a single channel. The value with global receptive field refers to the numerical distribution of feature images, and it can also be called global information. $F_{sq}(r_k)$ is implemented with global average pooling, which is calculated as

$$Z_k = F_{sq}(r_k) = \frac{1}{T \times 1} \sum_{i=1}^{W} \sum_{j=1}^{T} r_k(i, j), \tag{6}$$

where Z_k represents the weight of the compressed output; $r_k(i, j)$ is the value of row i and column j in the kth channel of the input features.

After retrieving the statistics from the squeeze operation, the excitation operation captures the correlation of the channels. Here, a simple gating mechanism is applied together with the Sigmod activation function, which is calculated as

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \tag{7}$$

where F_{ex} means the neural network parameter, σ denotes the Sigmod activation function, δ is the ReLU activation function, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in R^{\frac{C}{r} \times C}$ are the learnable parameters of two fully connected layers, and r means the reduction ratio. They are operated by the full connection layer, and then go through the ReLU activation function layer and Sigmoid activation function respectively to reduce and increase the dimension to restore the dimension to the original one. In this way, the weight information of feature images is obtained.

Finally, the weighting operations are optimized according to the excitation operation for each channel,

$$\widetilde{X}_k = F_{scale}(r_k, s_k) = s_k \cdot r_k, \tag{8}$$

where $\widetilde{X} = [\widetilde{X}_1, \widetilde{X}_2, \dots, \widetilde{X}_C]$ is the output matrix of the SE-block; $F_{scale}(r_k, s_k)$ is the dot product of the feature graph r_k and the weight s_k in the channel direction.

3.4. Attention-based aggregation

Connection attention mechanism used in AMSNet is covered in this section. The weight values of different types of features can be obtained through matrix operations following the connection method based on the attention mechanism. The weight value is used to redistribute the two types of features, and the relatively more important feature information is given a higher weight so that the feature has a larger contribution to the final recognition result. The specific calculation formula is as follows,

$$\beta_1 = \tanh(W_f h_1^T + b), \tag{9}$$

$$\beta_2 = \tanh(W_f h_2^T + b), \tag{10}$$

$$\alpha_i = \frac{\exp(\beta_i)}{\sum_{j=1}^{2} \exp(\beta_j)}, \tag{11}$$

where $i \in 1, 2$, W_f denotes the weight matrix, b is the bias, and h_1^T, h_2^T are the transpose of the outputs of the two parallel networks. Through the above operations, the weight values of the two features are determined, and the final representation can be obtained by multiplying and adding the corresponding output vectors,

$$h = \sum_{i=1}^{2} \alpha_i h_i. \tag{12}$$

3.5. Loss function

The cross-entropy loss function is the most common loss function used in classification tasks. Cross-entropy measures the distance between two different probability distributions. For a sample, the real tag and the model prediction tag can be represented by the cross-entropy function as,

$$l_{CE} = -\sum_{i=1}^{c} t_i \log(y_i). \tag{13}$$

where, $t_i = 1$ indicates that t_i belongs to the real category, and c is the dimension of the output distribution of the model.

The loss function of all samples can be obtained as:

$$l_{CE} = -\sum_{k=1}^{n}\sum_{i=1}^{c} t_{ki}\log(y_{ki}), \tag{14}$$

where t_{ki} denotes the probability that sample k belongs to category i, and y_{ki} is the probability that sample k is predicted by the model to belong to category i.

4. Experiments results and discussion

4.1. Datasets

IEMOCAP (Busso et al., 2008) is a real multi-modal, multi-speaker emotion database. The corpus, recorded by ten actors, consists of five conversations, each spoken by a male and a female. It contains approximately 12 h of audio data, including video, speech, motion capture, text transcription. This study only used audio data, including 10039 utterances. In the experiment, only the utterances with the following emotional labels were used: angry, happy, sad, and neutral. To balance the sample size for each emotion category, the emotions of happiness and excitement were combined into happiness. Finally, there were respectively 1103, 1636, 1084, and 1708 speech samples with the emotional labels of angry, happy, sad, and neutral, for a total of 5531 samples.

EmoDB is a German emotional language library recorded by the Technical University of Berlin. The dataset involves ten actors, five males and five females. The ten actors simulated seven emotions: neutral, angry, fear, happy, sad, disgust, and bored. The dataset consists of 233 sentences of male emotion statements and 302 sentences of female emotion statements, 535 sentences in total.

4.2. Feature extraction

Spectrogram Extraction: We calculate the average length of utterances in IEMOCAP database was 5s, so all utterances that are less than 5s were filled with silent fragments to 5s, while those larger than 5s are cut off to ensure data integrity to the greatest extent possible. Similarly, we calculate the average length of all utterances in EmoDB which valued 3s for cutting or filling. The length of the Hanning window was set to 800, and the sampling rate was set to 16 000 Hz. For each frame, a short-term Fourier transform with a length of 800 and a hop length of 400 was calculated. To simulate the nonlinear perception of sound by the human ear, the spectral map was mapped to the Mayer scale. The spectrogram was extracted as the input to the network. This study tried to use sampling windows of different lengths, but the results did not differ significantly.

LLDs Extraction: In the study, openSMILE was used to extract frame-level features. The extracted frame-level feature set contains features of 76 dimensions, including 34 dimensions of LLDs, 34 dimensions of LLDs first-order differences, 4 dimensions of pitches, 4 dimensions of pitch-dependent first-order differences. The dimension of the finally extracted frame-level feature set is $n \times q \times 75$, where n is the number of samples in the dataset, and q is the maximum time step.

4.3. Experimental setup

Implementation details: Due to the different lengths of the signals in the datasets, shorter signals need to be filled into the silent segment of the mean. The average signal length of the IEMOCAP dataset is 5s, and that of the EmoDB dataset is 3s. Therefore, this study used silent shards to fill voice shards that are less than a fixed second. The training set and validation set of the IEMOCAP and EmoDB all respectively account for 80% and 20% of the total samples, which is the same method applied in the references. 5-fold cross-validation was used to evaluate the effectiveness of the experiment. And the averaged accuracy was obtained for the model. The evaluation indexes used in this study are the same as those of domestic and foreign studies, which are weighted accuracy (WA) and unweighted accuracy (UA).

$$WA = \frac{\sum_{i=1}^{L} TP_i}{\sum_{i=1}^{L}(TP_i + FN_i)} \tag{15}$$

$$UA = \frac{1}{L}\sum_{i=1}^{L}\frac{TP_i}{TP_i + FN_i} \tag{16}$$

The input dimension of FRLM is $n{\times}q{\times}76$, where q denotes the length of the signal. The input is first passed to the dimension transpose layer. The transformed time series are then passed into the BLSTM block. The BLSTM block consists of the attention BLSTM layer followed by a dropout layer to prevent overfitting. The block outputs the spatial feature vector **X**, and it is finally processed by the global average pooling layer. The dimension of the final output is $u \times 128$, where u is the number of emotion categories in the dataset. For IEMOCAP, u equals 4; for EmoDB, u equals 7.

The input dimension of URLM is $n \times 1 \times 257 \times 1000$. The input is first fed into the SCNN block. The SCNN block is composed of three stacked time convolution blocks with filter sizes of 512, 256, and 128, and kernel sizes of 8, 5, and 3, respectively. The parameters were initialized with the normal distribution kaiming. For the SE-block, the reduction ratio r was set to 16. Each convolutional block is followed by a batch normalization layer, and the ReLU activation function is used. After the dense layer, the final output has a dimension of $u \times 128$.

The merge function is based on the attention mechanism, and it combines (H_1, H_2) of fixed network weights after pre-training of frame-level features and utterance-level features as the final features. After the fully connected layer with a dense value of 64, 8, and u, the predicted emotion categories are output.

Comparison with state-of-the-art methods: To further demonstrate the performance of AMSNet, the experimental results in the paper are compared with those in other papers.

FaceNet was adopted for SER tasks by Liu et al. (2021) for the first time. This work takes the spectrogram and the original waveform as input. Meanwhile, considering the difference between SER and face recognition, the triple loss is replaced with center loss.

HGFM (Zheng, Wang, & Jia, 2020) is a hierarchical grained feature model. It can capture features of different granularities and improve the sensitivity of the model to subtle clues. Meanwhile, a more complete representation of the hierarchical feature can be obtained from the raw speech signal.

The dual net (Chen & Huang, 2021) is composed of attention-based BLSTM, which uses linear interpolation and decimation to solve the problem of varying length of speech signals and inputs different features into the dual structure to obtain better effectiveness.

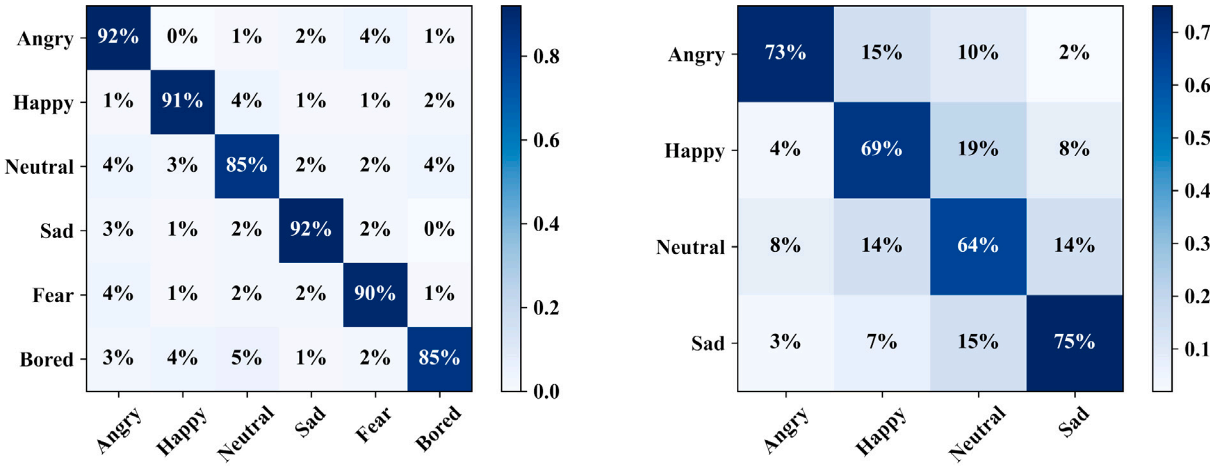
The system (Peng et al., 2020) combines auditory perception-based front-end and attention-based back-end, which extracts three-dimensional features by simulating the human auditory system.

The graph-attention mechanism is introduced into the gated recurrent unit network (Su, Chang, Lin, & Lee, 2020) to effectively reallocate the attention weight when processing complex acoustic coding characteristics of speech emotions, thus improving recognition accuracy.

XGBoost algorithm is introduced for feature selection in He and Ren (2021) to calculate the importance score of each feature and quantifies the impact of each feature on classification results. Meanwhile, CNN and BLSTM with an attention model are adopted to recognize speech emotion.

4.4. Experiment results analysis

The comparative experimental results based on the IEMOCAP English corpus and EmoDB Berlin corpus in Table 2 indicate that the AMSNet model proposed in this paper achieves the best recognition



(a)

(b)

Fig. 4. Confusion matrix of AMSNet on (a) The EmoDB dataset and (b) The IEMOCAP dataset, where each row presents the confusion of the ground-truth emotion during prediction.

Table 2
Results of the comparative experiments on IEMOCAP and EmoDB datasets.

Models	Dataset	WA (%)	UA (%)
FaceNet (Liu, Zhang et al., 2021)	IEMOCAP	68.96	69.53
HGFM (Xu, Xu, & Zou, 2020)	IEMOCAP	66.54	70.48
GA-GRU (Su et al., 2020)	IEMOCAP	62.27	63.80
3DCNN+ASRNN (Peng et al., 2020)	IEMOCAP	61.90	62.60
Dual Attention+BLSTM (Chen & Huang, 2021)	IEMOCAP	68.73	70.29
AMSNet	IEMOCAP	**69.22**	**70.51**
CNN (Issa, Demirci, & Yazici, 2020)	EMODB	86.10	86.94
XGBoost+CNN+BLSTM (He & Ren, 2021)	EMODB	86.87	86.03
AMSNet	EMODB	**88.34**	**88.56**

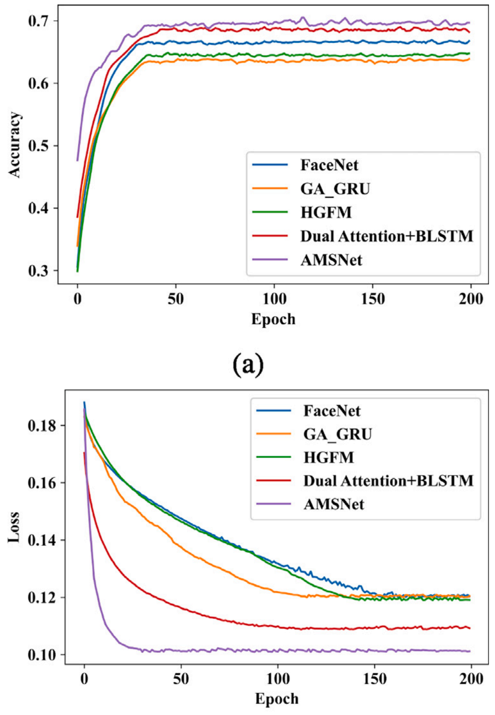
accuracy in WA and UA. Meanwhile, in the training process, the AMSNet also performs well and obtains a faster convergence rate.

It can be seen from Fig. 5 that the proposed AMSNet model has a higher recognition rate for emotions with stronger emotional colors, which is consistent with our cognition. Meanwhile, when artificially judging emotions, the recognition rate is also higher for more intense emotions such as anger and sadness. However, neutral emotions have a lower recognition rate due to their less obvious features. This phenomenon also occurs in the artificial recognition of speech emotion since the features of neutral emotions are not obvious enough, it is difficult to classify the corresponding speech signals (see Fig. 4).

4.5. Impact of various factors on model performance

In the construction of the model and the training process, many factors will affect the result. Optimizer selection and batch size setting are two factors affecting model performance. It is not difficult to find that batch size is set differently for different datasets in the experiments. As the specific experimental condition shown in Fig. 6, RMSprop can optimize the loss function and further speed up the convergence of the function, and using RMSprop optimizer is the best choice. For the IEMOCAP dataset with a large amount of data, the best choice of batch size is 64 or 128. For the EmoDB dataset with a small amount of data, a better recognition effect will be achieved if each epoch can cover more data, so the batch size is set to 256.

Learning rate is one of the most influential hyperparameters. The lower the learning rate is, the slower the loss function changes. However, if the learning set is too large, beyond the extreme value, the loss stops falling and oscillates repeatedly at a certain position. Experiments on this parameter were conducted in this study, and it was found that when the learning rate should be set within 5×10^{-5} and 10^{-4}.



(a)

(b)

Fig. 5. Training results of the comparative experiments on the IEMOCAP dataset with (a) ACC (%) value and (b) Loss value.

Table 3
The model performance and parameter sizes at different reduction ratios.

Ratio *r*	ACC (%)		Params
	IEMOCAP	EMODB	
2	70.34	87.94	23.1M
4	70.41	88.32	15.6M
8	70.53	88.59	12.5M
16	70.51	88.54	11.2M
32	70.39	88.42	10.8M

For URLM, the reduction ratio of the SE-block has some influence on the model performance. As shown in Table 3, during the experiment, it

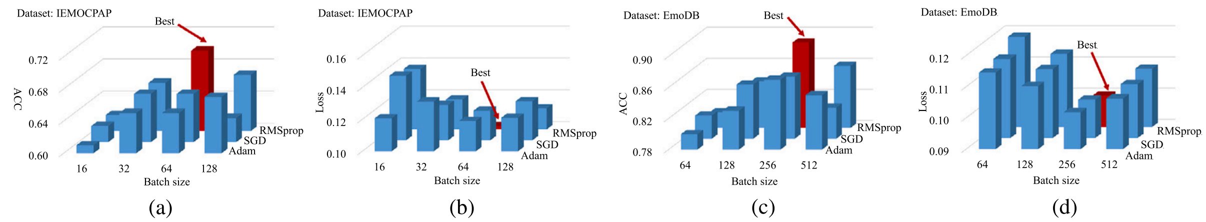


Fig. 6. The result obtained by AMSNet under different batch sizes and different optimizers. (a) and (b) show the impact of different optimizers and batch sizes on the ACC value and loss value on the IEMOCAP dataset. (c) and (d) show the impact of different optimizers and batch sizes on the ACC value and loss value on the EmoDB dataset.

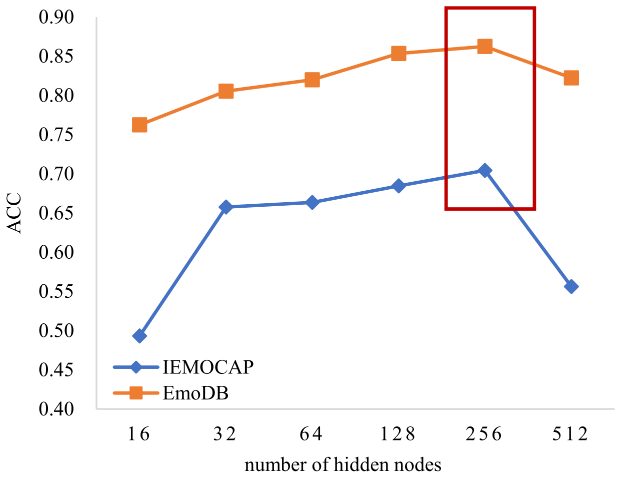


Fig. 7. Effect of different hidden sizes.

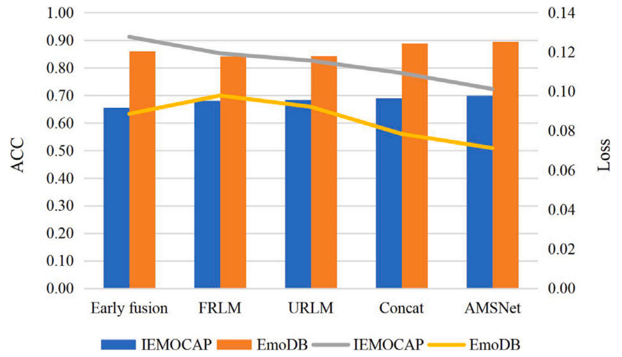


Fig. 8. Results of ablation experiments.

is appropriate to set the reduction ratio to 16 when both the accuracy and the number of parameters are considered.

Similarly, for FRLM, the hidden size of BLSTM is also an important factor affecting model performance. As shown in Fig. 7, when the hidden size is less than 128, the model does not perform well for SER; when hidden size is more than 512, the model is prone to overfitting and has a poor generalization ability. Therefore, hidden size can be set from 128 to 512 to obtain a better recognition effect.

4.6. Ablation experiments

To evaluate the effectiveness of the proposed AMSNet, five ablation experiments were conducted. Early fusion was selected to prove the effectiveness of fusion after feature training. URLM and FRLM methods were selected to prove the significance of fusing the two types of features. Besides, experiments were conducted on the concat method for fusion to prove that the attention mechanism can better adapt to SER tasks. In all the ablation experiments, the model was trained and evaluated on the IEMOCAP and EmoDB datasets. As shown in Fig. 8, in the training process, AMSNet achieves a positive convergence effect on various performance indicators.

Table 4
Results on the IEMOCAP dataset.

Models	WA (%)	UA (%)
Early fusion	65.69	66.36
FRLM	65.02	64.60
URLM	67.43	67.92
URLM+FRLM(Concat)	68.94	69.21
AMS NET	**69.22**	**70.51**

Table 5
Results on the EMODB dataset.

Models	WA (%)	UA (%)
Early fusion	85.63	86.32
FRLM	84.14	84.46
URLM	84.21	84.52
URLM+FRLM(Concat)	88.15	87.02
AMS NET	**88.34**	**88.56**

Meanwhile, it is found in Tables 4 and 5 that the results of AMSNet are improved significantly, indicating the effectiveness of all the structures of AMSNet.

5. Conclusion

In order to reduce the impact of the inherent defects of a single feature, AMSNet is proposed in this paper. Specifically, in the one hand, frame-level feature and utterance-level feature are fused to increase the complementarity between different features and combine the advantages of spatial and temporal features of signals. In the other hand, a fusion method based on the connection attention mechanism is proposed. Since different features have different contributions to the final recognition result, different weights are assigned to fully retain the advantages of feature fusion. Furthermore, extensive experiments have been conducted on two datasets to demonstrate the superiority of the proposed AMSNet. In the future, we will explore how to translate the method into real-time recognition.

CRediT authorship contribution statement

Zengzhao Chen: Supervision, Writing – review & editing. **Jiawen Li:** Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Hai Liu:** Conceptualization, Supervision, Project administration, Writing – review & editing. **Xuyang Wang:** Investigation, Conceptualization, Methodology, Investigation. **Hu Wang:** Conceptualization, Methodology, Investigation. **Qiuyu Zheng:** Methodology, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank the editor and anonymous reviewers for their valuable suggestions. This work has been supported by the National Natural Science Foundation of China (Grant No. 62077022, 61875068, 62211530433, 62177018, 62011530436, 62277041, 62005092, 62077020), the National Teacher Development Collaborative Innovation Experimental Base Construction Research Project of Central China Normal University (No. CCNUTEIII 2021-21), and the National Key R&D Program of China (2021YFC3340802). We would like to thank Dr. Tingting Liu and Prof. Zhaoli Zhang for assistance in advice.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440–1444.

Chen, Q., & Huang, G. (2021). A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. *Engineering Applications of Artificial Intelligence*, *102*, Article 104277.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Dolan, R. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191–1194.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2015). The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, *92*, 60–68.

Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE workshop on automatic speech recognition and understanding* (pp. 273–278). IEEE.

Guo, L., Wang, L., Dang, J., Liu, Z., & Guan, H. (2019). Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. *IEEE Access*, *7*, 75798–75809.

Guo, L., Wang, L., Dang, J., Zhang, L., & Guan, H. (2018). A feature fusion method based on extreme learning machine for speech emotion recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2666–2670). IEEE.

Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.

He, J., & Ren, L. (2021). Speech emotion recognition using XGBoost and CNN BLSTM with attention. In *2021 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computing, scalable computing & communications, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)* (pp. 154–159). IEEE.

Hou, M., Li, J., & Lu, G. (2020). A supervised non-negative matrix factorization model for speech emotion recognition. *Speech Communication*, *124*, 13–20.

Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, *59*, Article 101894.

Jiang, D.-N., & Cai, L.-H. (2004). Speech emotion classification with the combination of statistic features and temporal features. In *2004 IEEE international conference on multimedia and expo (ICME) (IEEE Cat. No.04TH8763), Vol. 3* (pp. 1967–1970). Vol.3.

Jiang, P., Fu, H., Tao, H., Lei, P., & Zhao, L. (2019). Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access*, *7*, 90368–90377.

Kumar, P., Jain, S., Raman, B., Roy, P. P., & Iwamura, M. (2021). End-to-end triplet loss based emotion embedding system for speech emotion recognition. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 8766–8773). IEEE.

Kumaran, U., Radha Rammohan, S., Nagarajan, S. M., & Prathik, A. (2021). Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. *International Journal of Speech Technology*, *24*(2), 303–314.

Kwon, S., et al. (2021). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications*, *167*, Article 114177.

Lanjewar, R. B., Mathurkar, S., & Patel, N. (2015). Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and K-nearest neighbor (K-NN) techniques. *Procedia Computer Science*, *49*, 50–57.

Li, D., Liu, J., Yang, Z., Sun, L., & Wang, Z. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, *173*, Article 114683.

Li, M., Yang, B., Levy, J., Stolcke, A., Rozgic, V., Matsoukas, S., et al. (2021). Contrastive unsupervised learning for speech emotion recognition. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6329–6333). IEEE.

Lin, Y.-L., & Wei, G. (2005). Speech emotion recognition based on HMM and SVM. In *2005 international conference on machine learning and cybernetics, Vol. 8* (pp. 4898–4901). IEEE.

Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K., & Wang, J. (2022). MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation. *IEEE Transactions on Multimedia*, *24*, 2449–2460.

Liu, S., Zhang, M., Fang, M., Zhao, J., Hou, K., & Hung, C.-C. (2021). Speech emotion recognition based on transfer learning from the FaceNet framework. *The Journal of the Acoustical Society of America*, *149*(2), 1338–1345.

Liu, H., Zheng, C., Li, D., Shen, X., Lin, K., Wang, J., et al. (2022). EDMF: Efficient deep matrix factorization with review feature learning for industrial recommender system. *IEEE Transactions on Industrial Informatics*, *18*(7), 4361–4371.

Luo, D., Zou, Y., & Huang, D. (2018). Investigation on joint representation learning for robust feature extraction in speech emotion recognition. In *Interspeech* (pp. 152–156).

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*(6), 9411–9457.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, *16*(8), 2203–2213.

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access*, *7*, 125868–125881.

Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2020). Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access*, *8*, 16560–16572.

Sahu, G. (2019). Multimodal speech emotion recognition and ambiguity resolution. arXiv preprint arXiv:1904.06022.

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech* (pp. 1089–1093).

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th annual conference of the international speech communication association, Lyon, France*.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M. (2011). Avec 2011–the first international audio/visual emotion challenge. In *International conference on affective computing and intelligent interaction* (pp. 415–424). Springer.

Su, B.-H., Chang, C.-M., Lin, Y.-S., & Lee, C.-C. (2020). Improving speech emotion recognition using graph attentive bi-directional gated recurrent unit network. In *INTERSPEECH* (pp. 506–510).

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., et al. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 3–10).

Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L. (2015). Speech emotion recognition using Fourier parameters. *IEEE Transactions on Affective Computing*, *6*(1), 69–75.

Wang, K., Wang, C., Wang, Y., Luo, W., Zhan, P., Hu, Y., et al. (2021). Time series classification via enhanced temporal representation learning. In *2021 IEEE 6th international conference on big data analytics (ICBDA)* (pp. 188–192). IEEE.

Wu, X., Cao, Y., Lu, H., Liu, S., Wang, D., Wu, Z., et al. (2021). Speech emotion recognition using sequential capsule networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3280–3291.

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(11), 1675–1685.

Xu, Y., Xu, H., & Zou, J. (2020). HGFM: A hierarchical grained and feature model for acoustic emotion recognition. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6499–6503). IEEE.

Xu, M., Zhang, F., & Zhang, W. (2021). Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access*, *9*, 74539–74549.

Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y. (2018). Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 1771–1775). IEEE.

Zhang, R., Wu, H., Li, W., Jiang, D., Zou, W., & Li, X. (2021). Transformer based unsupervised pre-training for acoustic representation learning. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6933–6937). IEEE.

Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., Tao, J., et al. (2021). Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition. *Neural Networks*, *141*, 52–60.

Zhao, J., Mao, X., & Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep CNN. *IET Signal Processing*, *12*(6), 713–721.

Zheng, C., Wang, C., & Jia, N. (2020). An ensemble model for multi-level speech emotion recognition. *Applied Sciences*, *10*(1), 205.