Applied Soft Computing 130 (2022) 109648

Contents lists available at ScienceDirect

Applied Soft Computing

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

Multi-type features separating fusion learning for Speech Emotion Recognition

Xinlei Xu [a,b], Dongdong Li [b,*], Yijun Zhou [b], Zhe Wang [a,b,**]

[a] Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science & Technology, Shanghai, 200237, China
[b] Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China

ARTICLE INFO

Article history:
Received 5 March 2022
Received in revised form 22 July 2022
Accepted 13 September 2022
Available online 23 September 2022

Keywords:
Speech emotion recognition
Hybrid feature selection
Feature-level fusion
Speaker-independent

ABSTRACT

Speech Emotion Recognition (SER) is a challengeable task to improve human–computer interaction. Speech data have different representations, and choosing the appropriate features to express the emotion behind the speech is difficult. The human brain can comprehensively judge the same thing in different dimensional representations to obtain the final result. Inspired by this, we believe that it is reasonable to have complementary advantages between different representations of speech data. Therefore, a Hybrid Deep Learning with Multi-type features Model (HD-MFM) is proposed to integrate the acoustic, temporal and image information of speech. Specifically, we utilize Convolutional Neural Network (CNN) to extract image information from the spectrogram of speech. Deep Neural Network (DNN) is used for extracting the acoustic information from the statistic features of speech. Then, Long Short-Term Memory (LSTM) is chosen to extract the temporal information from the Mel-Frequency Cepstral Coefficients (MFCC) of speech. Finally, three different types of speech features are concatenated together to get a richer emotion representation with better discriminative property. Considering that different fusion strategies affect the relationship between features, we consider two fusion strategies in this paper named separating and merging. To evaluate the feasibility and effectiveness of the proposed HD-MFM, we perform extensive experiments on EMO-DB and IEMOCAP of SER. The experimental results show that the separating method has more significant advantages in feature complementarity. The proposed HD-MFM obtains 91.25% and 72.02% results on EMO-DB and IEMOCAP. The obtained results indicate the proposed HD-MFM can make full use of the effective complementary feature representations by separating strategy to further enhance the performance of SER.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In general applications such as psychology [1–3], emotion is a part of the human mental state, which represents the true feelings of human beings about a certain thing. Emotional recognition can be considered multi-dimensional. For example, human expressions, eyes, speech and electroencephalogram all can be used as the basis of emotion recognition. In recent years, with the increasing demand for human–computer interaction, speech emotion recognition (SER) has attracted more and more attention [4–6]. SER can be widely used in many fields, such as natural human–machine interaction, disease diagnosis [7], fatigue detection [8], public security etc. A traditional SER system consists of two parts: one is to explore the appropriate emotion representative features [9], the other is to establish a classifier adapted to emotion recognition [10].

The features which are used to represent the speech emotion are usually divided into two types, the handcrafted Low-level descriptors (LLDs) and the high-level statistic features (HSFs) [11]. Researchers split speech signals into small segments called frames, and then process these frames separately. The features extracted from these frames are called LLDs. Some commonly used LLDs are fundamental frequency (F0), energy, formants, spectrogram, Mel-Frequency Cepstral Coefficients (MFCC), shimmer, jitter, etc. On the other hand, HSFs are calculated as statistics of all speech features extracted from an utterance, which means they are calculated as statistics of LLDs. For classifiers, a variety of models have been applied for classification based on those speech features. Gaussian Mixture Models (GMM) [12–14] and the Hidden Markov Model (HMM) [15–17] are the most commonly used in SER. Daneshfar [18] employed modified Quantum-behaved

* Corresponding author.
** Corresponding author at: Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China.
E-mail addresses: ldd@ecust.edu.cn (D. Li), wangzhe@ecust.edu.cn (Z. Wang).

https://doi.org/10.1016/j.asoc.2022.109648
1568-4946/© 2022 Elsevier B.V. All rights reserved.

Particle Swarm Optimization (QPSO) algorithm to estimate both the optimal projection matrix for feature-vector dimension reduction and GMM classifier parameters. Other commonly used classifiers are Support Vector Machine (SVM) [19,20], k-nearest neighbor (KNN) [21,22], Decision Trees [23] etc.

Nowadays, with the development of Deep Learning (DL) [24–26], it has been widely applied in various fields included SER [27]. Deep Neural Network (DNN) is composed of multiple perceptrons. DNN has a strong performance in learning plane-independent structural information, so it is often used to learn HSFs in SER [28,29]. Convolutional Neural Network (CNN) [30] is also a feed forward neural network. The main characteristic of CNN is that its intrinsic convolution structure makes it have a better ability in image information learning. Much work has been done to study its performance in learning emotional expression from spectrogram [31,32]. Kwon [33] proposed a one-dimensional CNN for real-time SER system. Lee [34] proposed the Fusion-ConvBERT exploit available information from given speech signals to the maximum extent possible. Recurrent Neural Network (RNN) is another popular deep model in the field of SER. The advantage of RNN is that it introduces the memory cell into the model, which makes it has a strong performance on dealing with the temporal information. Kumaran [35] proposed a Deep Convolutional–Recurrent Neural Network (Deep C-RNN) to classify the effectiveness of learning emotion variations in the classification stage. Long Short-Term Memory (LSTM) [36–39] is an improved model of RNN, which reduces the problem of gradient disappearance and can memorize information for a longer time. Many speech features extracted from frames have rich temporal information, so the LSTM is often used in SER [32,40]. Pend [41] proposed a parallel LSTM architecture to model the multi-temporal dependencies of the multi-resolution modulation-filtered cochleagram. The recent speech emotion recognition attempts to combine speech and information in this text to improve the classification performance of SER. Peng [41] proposed a simple yet efficient neural network architecture to exploit both acoustic and lexical information from speech. Fan [42] proposed an adaptive domain-aware representation learning that leverages the domain knowledge to extract domain aware features. Li [43] proposed a machine learning framework to obtain speech emotion representations by limiting the effect of speaker variability in the speech signals.

Researchers have found that speech features can be expressed in higher dimensions through DL, which cannot be achieved by traditional methods. At the same time, the different deep models are distinguished from each other, so the high-dimensional representations of speech features obtained from these different models will have some complementary information. In order to achieve and learn the complementarities behind the speech features, many researchers have launched several works in speech features combination based on deep learning. In [44], the author aimed to enrich the features and combine two types of features, one is extracted by a CNN and the other is HSFs. In [45], the author proposed an HSF-CRNN SER system, which replaced the CNN with CRNN, to learn the joint representation of CRNN-learned features and HSFs through the neural network. In [46], an Attention-BLSTM-RNNs method was proposed, which used a fully convolutional network (FCN) to extract high-level spatial features and a BLSTM with an attention network to extract high-level temporal features, then concatenate them into a DNN to predict the final emotion. These works enable the neural networks to obtain the interdependencies between different emotional expression spaces. In addition, these methods above are effective in learning speech features, so it can be used as a suitable candidate to further improve the models.

Inspired by the ability of the human brain to make judgments through multi-dimensional information, we explore the fusion strategy of speech information of different representations to improve the ability of SER. Different from the above works, this paper aims to utilize the characteristics of different neural networks to learn the high-level representations of speech features in three different dimensions: image, temporal and statistic. Hence, the Hybrid Deep Learning with Multi-type features Model (HD-MFM) integrated three distinctive neural networks is proposed to discriminate four emotional states: angry, happy, neutral and sad. Related studies show that CNN, DNN and LSTM have excellent capabilities in extracting different image information. Therefore, we use CNN to obtain image information of the spectrogram. DNN is utilized to search statistical representation from HSFs, and LSTM is used to learn temporal information from MFCCs. After that, a high-level representation of speech features from three different dimensions is obtained. Then, we further explore two different feature fusion methods named separating and merging. These three types of features with different information are joined together to obtain a new effective complementary speech feature vector. HD-MFM utilizes the complementary speech feature vector to recognize the final emotion.

The main contributions of this paper are summarized as follows:

- This paper proposes a SER system called HD-MFM, which can obtain different information from three dimensions of speech and get an effective complementary speech feature representation.
- This paper considers two different feature combination strategies, named separating and merging. The experimental results show that the separating can achieve the best performance than merging when fusing different dimensions speech features.
- This paper proves experimentally that HD-MFM can learn emotional information from speech features very well, and get the best performance compare with other works in SER which also fuse features through multiple models.

The rest of the paper is structured as follows: In Section 2, the proposed framework is explained. Then, the database and the method for extracting the speech features are presented in Section 3. After that, the experiment setup and the results are discussed in Section 4. Finally, the conclusions of this work are provided in Section 5.

2. Proposed method

In this work, we introduce the proposed HD-MFM, which can detect the complementarities among the different dimension features by utilizing the characteristics of different deep learning models. The overall structure of the HD-MFM is shown in Fig. 1. It consists of three blocks, Spectrogram-CNN block, HSF-DNN block and MFCC-LSTM block. The training process of our method can be summarized in three steps. First of all, we extract three different types of features from the raw wave file, including spectrogram, HSFs and MFCC. Then three different types of features are utilized as the input of three different deep learning models to train the features respectively and extract high-dimensional feature representations. Finally, we combine these three representations together and use two fully-connected layers to test the effect of this fusion feature. In the following sections, we first introduce the three modules and their respective applied speech features. Then the two fusion strategies are introduced and compared in detail, and the separation strategy is finally selected.

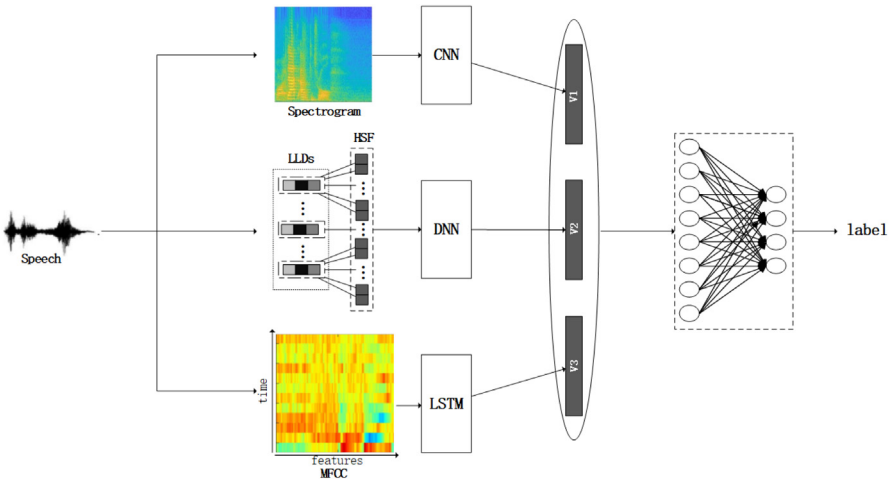


Fig. 1. The flow chart of the HD-MFM.

2.1. Spectrogram-CNN block

In this block, we use the spectrogram to learn the image information of speech. The spectrogram is usually obtained by processing the received time-domain signal [47]. The spectrogram can represent most of the speech information, frequency, energy intensity, peak value and also the length of a speech in the form of an image. The abscissa of the spectrogram is time, and the ordinate is frequency, and the coordinate point is speech data energy. Because it uses a two-dimensional plane to express three-dimensional information, the energy value is expressed by color. The darker the color, the stronger the voice energy of the point is.

In order to get the spectrogram, the raw signal $S = [s(1), s(2), \ldots, s(T)] \in \Re^{1\times T}$,where T is the length of the signal, should be pre-processed by pre-emphasis, frame division and windowing at first. After that, the raw signal S is turned into segment frames $F = [f(1), f(2), \ldots, f(K)] \in \Re^{K\times M}$, where K is the frame number, M is the frame length. Then an N-point Fast Fourier Transformation (FFT) is used on each frame to calculate the frequency spectrum. This transformation is also called Short-Time Fourier-Transform (STFT), where N is typically 256 or 512. Then the spectrogram is completed by using the following equation:

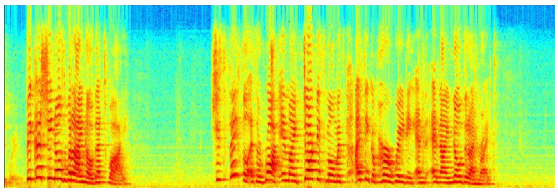
$$P_i = \frac{|FFT(f_i)|^2}{N} \tag{1}$$

where, f_i is the ith frame of signal S. So after FFT transformation, the segment frames F is transformed into the Spectrogram $P = [p(1), p(2), \ldots, p(K)] \in \Re^{K\times Q}$, where $Q = \frac{N}{2}$. Because the K of each speech signal is different, we will get matrices of different sizes in time dimension after FFT transformation. In this work, to consider spectrograms as image features, the bilinear interpolation algorithm is used to compress spectrograms to obtain images of uniform dimension size:

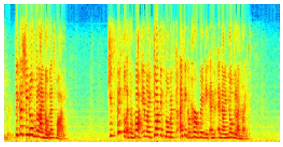
$$\begin{aligned} f(i+x, j+y) = {} & xyf(i+1, j+1) \\ & +(1-x)(1-y)f(i,j) \\ & +(1-x)yf(i, j+1) \\ & +x(1-y)f(i+1, j) \end{aligned} \tag{2}$$

where f(i,j) denotes the pixel values corresponding to the coordinates (i, j) of the original spectrogram. As can be seen from Fig. 2, the new image after transformation is slightly different from the original image, and retains a lot of detail information about the original image.

The new spectrogram $P' \in \Re^{K'\times Q}$ is fed into a CNN, which is comprised with four convolutional layers and two fully-connected layers. The recent works [30] show that the CNN can effectively extract image information and complete image classification tasks because CNN can fully consider the spatial information of the image through the filter. Therefore, for the spectrogram of speech data, we want to extract the speech emotion classification information contained in the spectrogram through CNN [48,49]. The diagram of the Spectrogram-CNN Block is shown in Fig. 3. The input feature is spectrogram. The dimension of the spectrogram is $256\times K$, where K depends on the length of the utterance. Hence the bilinear interpolation is used to transform it into 256×256. The filter of the convolution layer is 100, the kernels size is 3×3, the stride size is 2. The kernels size of pooling layer is 2×2, the stride size is 1. The two fully-connected layers contain 2048 and 1024 hidden units respectively. Each CNN and fully-connected layer contain BatchNorm and ReLU. The dropout layer is used after each fully-connected layer. The probability of dropout abandonment is 0.5. The output shape of Spectrogram-CNN block is transformed into a 1×1024 vector.



(a) Original Spectrogram



(b) Transformed Spectrogram

Fig. 2. Examples of the original image and the converted image: The file name is Ses01F_script01_2_F008.wav, and the emotion is angry.

The most distinguishing characteristic of the convolution layer is the local association brought by sliding window and the mechanism of parameter sharing. The convolution network used in this paper can be expressed by the following equation:

$$h_{jk}^{l} = \sum_{j}\sum_{k} w_{jk}^{l} x_{jk}^{l-1} + b^{l} \tag{3}$$

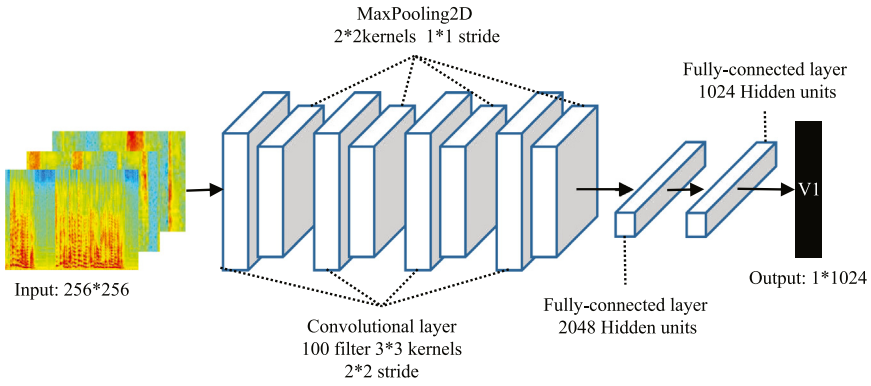


Fig. 3. The diagram of the Spectrogram-CNN Block: It consists of four convolution layers and the combination of maximum pooling layers, and then, the feature is transformed into vector V_1 through full connection layer.

$$x_{jk}^{l} = \sigma(h_{jk}^{l}) \tag{4}$$

where w_{jk}^{l} denotes the convolution kernel weights of the area (j, k) in the lth layer, and according to the number of convolution kernels n, there can be multiple different weights w, that is, in this equation, $w \in \Re^{n*j*k}$. x_{jk}^{l-1} means the value of the (j, k) region at the $(l-1)$ layer, h_{jk}^{l} denotes the preactivation output of the jth unit in the lth layer, b^{l} is a bias added to each convolution kernel in the lth layer and σ is the non-linear activation function. The activation function used in this paper is the Rectified Linear Unit (ReLU) due to its computational simplicity and faster learning convergence [50], and it can be represented as follows:

$$\sigma(x) = \begin{cases} x & x \geq 0 \\ 0 & otherwise \end{cases} \tag{5}$$

The function of pooling layer is to remove the irrelevant information of image, reduce the dimensions of features and prevent the over-fitting of neural network to a certain extent. The common used pooling methods are average pooling and max pooling, and the max pooling is used most. The features produced by max-pooling layer can be expressed as follows:

$$a_{ij}^{l} = max(a_{mn}^{l-1}) \tag{6}$$

where a_{ij}^{l} represents the output of the pooling layer with index i and j, a_{mn}^{l-1} represents the region covered by the pool core corresponding to $(l-1)$th layer with index m and n. Therefore, maxpooling achieves feature dimensionality reduction and nonlinear transformation by outputting the maximum of each region. Through the CNN, the spectrogram P will be changed into a feature vector $V_1 \in \Re^{1\times n_1}$ by two fully-connected layers, where n_1 is the unit number of the last fully-connected layer, carried the representation of image dimension.

2.2. HSF-DNN block

As mentioned above, HSF is a feature set of statistical algorithms for various LLDs. Its rich statistical information on each various prosodic, spectral and voice quality features make it best reflect the overall information of speech. The advantage of selecting the HSFs is that its less feature dimensions with lower classification time and higher classification accuracy. There are many commonly used HSF sets, such as GeMAPS [51], GeMAPS with extended(eGeMAPS), Interspeech′09 and Interspeech′10.

There are 62 features in GeMAPS feature set, which are obtained from 18 LLD features by statistical calculation. Table 1 shows the names of 18 LLDs in GeMAPS. eGeMAPS is an extension version of GeMAPS. On the basis of 18 LLDs, cepstrum features are added, including MFCC 1-4, Spectral flu and Formant 2, 3 bandwidths. A total of 25 LLD features, and 88 feature sets were obtained by statistical calculation. Interspeech′09 and Interspeech′10 are HSF feature set proposed on emotional challenge of INTERSPEECH in 2009 and 2010, respectively. Interspeech′09 chooses 16 LLD features, which is showed in Table 2, and their corresponding first-order derivative features. So that a total of 32 acoustic features are used as basic features, and 12 statistical functions are used to calculate them respectively. Thus, a feature set containing 384 speech features is obtained. Interspeech′10 uses 38 LLD features, which are shown in Table 2, and their corresponding first-order derivative features as the basic features, and uses 21 feature statistical functions to carry out feature statistics on 68 LLDs, which without the 4 LLDs of baseline frequency and their derivatives, obtaining 1428 features; Then 19 feature statistical functions are used to count these eight LLDs, and 152 features are obtained. In addition, two additional features pitch interval and total duration are added. Thus, the total feature number of the Interspeech′10 is 1582. In this paper, after extensive experimental verification, we finally choose Interspeech′10 due to its better HSF performance.

In order to fully acquire and utilize the speech information contained in HSFs, a two fully-connected layers of DNN is used to learn it, which is shown in Fig. 4. Both the hidden units of fully-connected layers are 1024. BatchNorm and ReLU are also

Table 1
LLD features contained in GeMAPS.

Type	Number	Name
Frequency related	6	Pitch, Jitter, Formant 1, 2, and 3 frequency bandwidth of first formant.
Energy/Amplitude related	3	Shimmer, Loudness Harmonics-to-noise ratio (HNR)
Spectral (balance) related	9	Alpha Ratio, Hammarberg Index Spectral Slope 0–500 Hz & 500–1500 Hz Formant 1, 2, and 3 relative energy Harmonic difference H1-H2 & H1-A3

Table 2
LLD features contained in Interspeech′09 and Interspeech′10.

Dataset	Number	Name
Interspeech′09	16	Zero Crossing Rate, Pitch Frequency Root Mean Square (RMS) Frame Energy Harmonics-to-Noise Ratio, MFCC[1–12]
Interspeech′10	38	PCM loudness, F0 envelop Log Mel freq. band[0–7] Voicing probability, MFCC[0–14] F0final, LSP frequency[0–7] jitterLocal, jitterDDP, shimmerLocal

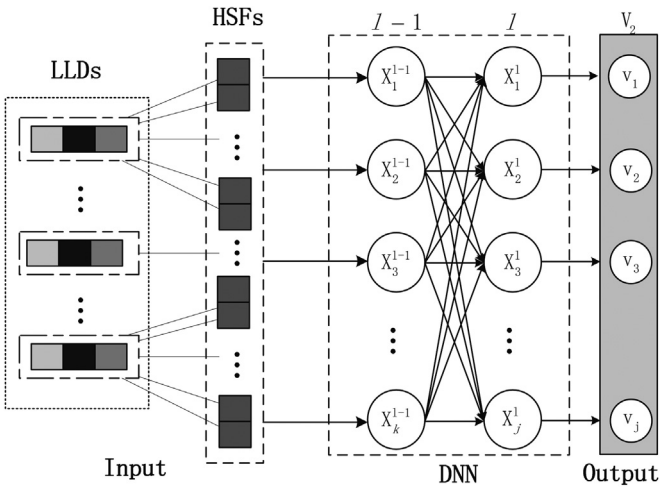


Fig. 4. The diagram of the HSF-DNN Block.

applied, and the dropout layer is used after each fully-connected layer with probability 0.5. The reason for choosing the DNN is that some researches [52] have demonstrated that DNN can effectively extract discriminative features that approximate the non-linear dependencies between features in the original set. Since the statistical HSF is a single-dimensional feature vector, it does not contain spatial information. DNN has a strong performance in learning plane-independent structural information, so it is often used to learn HSFs in SER. DNN can be modeled by iteration of the following two formulas:

$$h_j^l = \sum_k w_{jk}^l x_k^{l-1} + b_j^l \tag{7}$$

$$x_j^l = \sigma(h_j^l) \tag{8}$$

where h_j^l denotes the preactivation output of the jth unit in the lth layer, w_{jk}^l denotes the connecting weight of the jth unit in the lth layer from the kth unit in the $(l-1)$th layer, x_k^{l-1} denotes the output of the kth unit in the $(l-1)$th layer, b_j^l is a bias added to the jth unit in the lth layer, and σ is the non-linear activation function Relu, which is presented in Eqs. (5). In this block, the workflow is as follows:

- Given an utterance $S = [s(1), s(2), \ldots, s(T)]$, and segmented it into frames $F = [f(1), f(2), \ldots, f(K)]$.
- For each frame F, extracted the LLDs, $L = [l(1), l(2), \ldots, l(\ell)] \in \Re^{K\times\ell}$, where ℓ represents the number of LLDs to be extracted, and computed the HSFs, $H(l) = [H_1(l), H_2(l), \ldots, H_\eta(l)] \in \Re^{1\times\eta\ell}$, where η is the number of statistical functions to be used.
- Put the HSFs into the DNN, and used the output value as the statistic feature vector $V_2 \in \Re^{1\times n_2}$, where n_2 is the unit number of the last DNN layer. In this paper, the output shape of HSF-DNN block is transformed into a 1×1024 vector.

2.3. MFCC-LSTM block

This block is to obtain the representation of speech in time dimension. The Cepstrum-based speech features are often considered to have abundant temporal information. MFCC is a common cepstrum feature in SER [53]. The principle of this parameter is based on the characteristics of human ear auditory mechanism. The reason to choose MFCC as our feature is that it has temporal information and lower feature dimensions. The MFCC is obtained by Discrete Cosine Transform (DCT), $Melcep = [M(1), M(2), \ldots, M(c)] \in \Re^{K\times c}$, where c means the cepstral coefficients, and often the c is 13 [54–57].

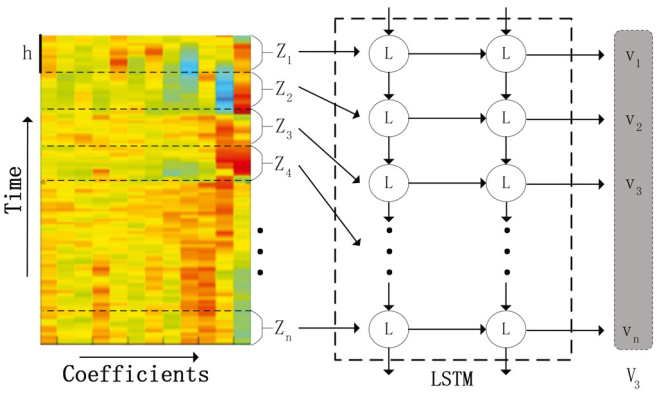


Fig. 5. The diagram of the MFCC-LSTM Block.

A two-layer LSTM model [58,59] is used in this block. LSTM is a recurrent neural network, which is designed specifically for learning long-term dependencies from sequences. Since MFCC contains rich temporal information, LSTM can extract the SER features related to temporal information. An LSTM cell can be described using Eqs. (9)–(14):

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{9}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{10}$$

$$\widetilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{11}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{12}$$

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{13}$$

$$h_t = O_t + \tanh(C_t) \tag{14}$$

where f_t, i_t, C_t and O_t denotes four main components of the LSTM, forget gate, input gate, cell state and output gate. W_f, W_i and W_o denotes the weights of forget gate, input gate and output gate. b_f, b_i, b_c and b_o denotes the bias of forget gate, input gate, cell state and output gate. $\widetilde{C}_t$ is a new candidate value that can be added to the cell state, σ is a sigmoid function and tanh is a hyperbolic tangent (tanh) function.

The diagram of the MFCC-LSTM Block is shown in Fig. 5. The first LSTM layer has 512 hidden units, and the second LSTM layer owns 256 hidden units. The 39MFCCs is used in this work. After extensive experimental verification, the time steps of the LSTM are 60, which means only the first 60 frames of information will be selected for each utterance. Hence, the 60 frames contain $59 \times 0.01 + 0.025 = 0.615$ s of speech information. Research has shown that a speech segment longer than 250 ms can encode sufficient emotion information [60]. Thus, through the LSTM, a feature vector $V_3 \in \Re^{1\times n_3}$, where n_3 is the unit number of the last LSTM layer, with sufficient temporal information can be obtained. In addition, considering the selection of different time intervals h is also one of the factors affecting emotional recognition [46]. For different time intervals, corresponding comparative experiments have been done in this paper, and a better time interval of emotional classification has been selected as the value of this block. The output shape of MFCC-LSTM block is transformed into a 1×256 vector.

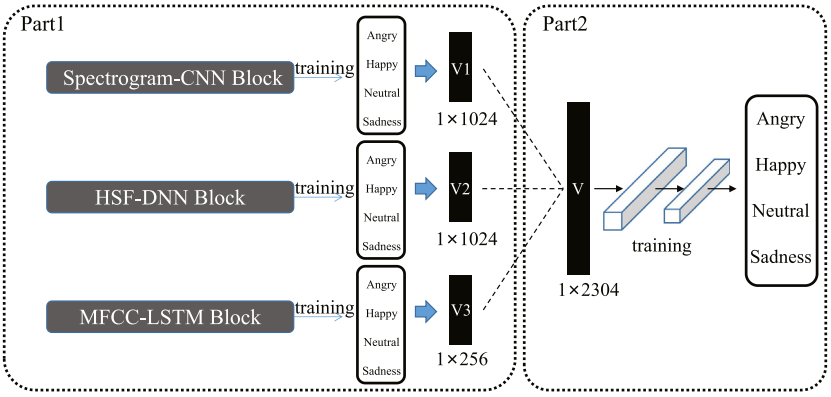


Fig. 6. The diagram of the separating. In the part1, the three blocks independently train the model of speech emotion recognition, and the corresponding feature vectors $V_1 \in \Re^{1\times n1}$, $V_2 \in \Re^{1\times n2}$, $V_3 \in \Re^{1\times n3}$ of the three blocks are obtained. In the part2, the three feature vectors are concatenated into a larger feature vector V, and then the speech emotion recognition classification is performed through another neural network with two fully connected layers.

2.4. The fusion strategy of HD-MFM

As stated before, the Spectrogram-CNN Block is used to capture the speech representation in image-dimension. The HSF-DNN Block gets the speech representation in statistic-dimension. The MFCC-LSTM Block captures the speech representation in time-dimension. These three types of features describe the emotional state of speech from different aspects and are mapped in respective feature spaces. Therefore, there is complementarity among these three kinds of features. Since the features of the three blocks contain speech information of different dimensions, this paper attempts to complement the advantages of three dimensions features in a certain way. We propose the overall framework HD-MFM to fuse the above three speech information of different dimensions. This paper assumes two data fusion methods in HD-MFM to obtain the best model, namely separating and merging.

Among them, separating method divides the overall model into two parts, as shown in Fig. 6. In the part1, three classification models are independently trained with the corresponding speech representations to complete emotion classification. After training three neural networks, the three feature vectors $V_1 \in \Re^{1\times n1}$, $V_2 \in \Re^{1\times n2}$, $V_3 \in \Re^{1\times n3}$ are then concatenated together into a vector $V = [V_1, V_2, V_3] \in \Re^{1\times n}$, where $n = n_1 + n_2 + n_3$. The output V_1 of Spectrogram-CNN block is 1×1024. The output V_2 of HSF-DNN block is 1×1024. The output V_3 of MFCC-LSTM block is 1×256. The dimension of the joint feature vector V is 1×2304. In the part2, we utilize the V as input to train another network with two fully-connected layers and a softmax layer with 4 nodes to complete speech emotion recognition. The two fully-connected layers have 2048 and 1024 nodes respectively. Each fully-connected layer contains BatchNorm and ReLU. The dropout layer is used after each fully-connected layer, and the probability is set 0.5. V can be mapped to a new feature space Y through the network. Then, one softmax layer is applied to complete the SER task.

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{J} e^{z_i}} \tag{15}$$

$$\hat{y} = arg\max_i p_i \tag{16}$$

where z_i is the input of softmax, J is the number of output classes, p_i is the probabilities of each emotion classes, $\hat{y}$ is the predicted class label. The advantage of separating can keep the independent features of each blocks from the influence of other blocks. The combining new feature vector can ensure the fusion of the information of the three blocks. In general, all blocks make up for the lack of information with each other, and further enhance the SER capabilities.

Different from separating, merging concatenates three blocks to an overall network for speech emotion recognition training. There is no separate parts, as shown in Fig. 7. Constrain the overall model through a single cross-entropy classification loss to train from scratch. The advantage of merging is that each block can interact to update the gradient through the global loss in the training process. The merging strategy obtains an end-to-end model. The input of the model is speech data in three representations, and finally the output of the softmax layer is used as the final result. Finally, a classifier that satisfies the three blocks at the same time is obtained.

In our final HD-MFM, separating is utilized through extensive experiments. Separating can ensure sufficient training of each block so that blocks do not interfere with each other. Each trained block model fully considers the characteristics of different types of speech data. Besides, another network is used to further complete the speech emotion classification without losing the original feature information of each block.

3. Database and feature extraction

3.1. Database

The emotional database used in the present work is the publicly available Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus [61] and Berlin database (EMO-DB) [62].

IEMOCAP database consists of five sessions from 10 professional speakers (five male and five female) in a professional studio recorded about 12 h of audio and video content, including improvisation and pre-script. Utterances are labeled emotion by majority voting. For all ten emotions, five of them are selected to participate in the experiment: angry(1103), excitement(1040), happy(595), neutral(1708), sad(1084). Because the expressions of happiness and excitement are similar, we combine these two emotions into one category.

EMO-DB consists of 10 professional actors (5 males and 5 females) in a noise-and-echo-less soundproof room uttered ten identical German short sentences with a total 535 utterances of seven kinds of emotions, anger (127), anxiety fear (69), boredom (81), disgust (46), happiness (71), neutral (79), and sadness (62). Corresponding to IEMOCAP, anger, happiness neutral and sadness were selected for experiment.

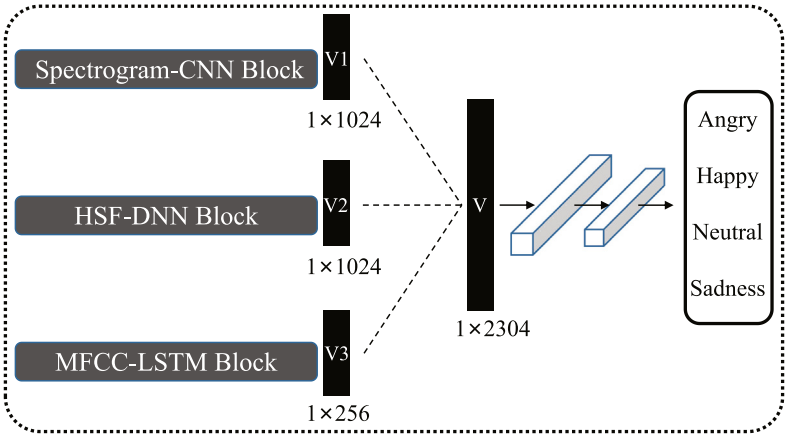


Fig. 7. The diagram of the merging. The three blocks and final emotion classification are trained in one network.

3.2. Feature extraction

The features used in this experiment is as follows: (1) Spectrogram, (2) 39MFCCs and (3) HSFs. To get these features, the speech signal is converted into a short-time frames using a 25-ms frame size and 10-ms frame sliding. After that, a Hamming window is applied to each frame. Then, the spectrogram will be obtained by the FFT which is showed in Eqs. (1). Next step is to apply triangular filters, typically 40 filters. Calculate the Mel-frequency distribution of these 40 filters, and then convert the Mel-frequency into the actual frequency. In order to reduce the high correlation of the filter bank coefficients, the DCT is applied to decorrelate the filter bank coefficients and yield a compressed representation of the filter banks. Thus, the MFCC is extracted. In order to make MFCC more recognizable, the first 13 order coefficients are selected and their corresponding first and second derivatives respectively are also taken. So, in this work, MFCC is in 39 dimensions. The HSFs are calculated by statistical function on LLDs, such as mean, max, min, kurtosis, and skewness. HSFs can be extracted via the freely available *openSMILE* tool [63].

4. Experiment set and result

4.1. Experiment setup

Through this section, we mainly explain the set up for the proposed HD-MFM. The details of the structure is shown in Table 3. In the Spectrogram+CNN block, a deep CNN network [48,49] is used for combining with four convolution layers and maxpooling2D layers and two fully-connected layers. The kernel size of the convolution layer is 3×3, the stride size is 2, and the kernel size of pooling layer is 2×2, the stride size is 1. Each CNN and fully connected layer contain BatchNorm and ReLU. The dropout layer is used after each fully-connected layer with probability 0.5. The output shape of this block is transformed into a 1×1024 vector. The input feature is spectrogram. The bilinear interpolation is used to transform spectrogram into 256×256. In the HSF-DNN Block, a DNN network with two fully-connected layers is used. Each of the layer has 1024 units, and comprises with a BatchNorm, ReLU and a dropout layer. The dropout probability is also set as 0.5. The interspeech′10 dataset is utilized as the HSF, due to the best performance in our experiment. To get the HSF features, the *openSMILE*[63] toolkit is used. The position features in the interspeech′10 such as 'pcm_loudness_sma_maxPos', 'pcm_loudness_sma_minPos' are changed into the position rate, e.g. one of the utterance's maxPos is 95, the minPos is 36 and the length of the utterance is 1.8, so the maxPosrate equals $95/180 = 0.5277$, the minPosrate equals 0.2. In addition, another two features called 'F0final_Turn_numOnsets' and 'F0final_Turn_duration' are abandoned in this work. In the MFCC-LSTM Block, a two-layer LSTM network [58,59] is used in this block. The 39MFCCs is used in this work. The time steps of the LSTM are 60, which means only the first 60 frames of information will be selected for each utterance. Hence, the 60 frames contain $59 \times 0.01 + 0.025 = 0.615$ s of speech information. Research has shown that a speech segment longer than 250 ms can encode sufficient emotion information [60]. The dimension of the joint feature vector V is 1×2304. The joint learning fully-connected layers have 2048 and 1024 nodes respectively and the softmax layer has 4 nodes. Before input into each block, speech features are normalized by the z-score normalization function.

$$\hat{x} = \frac{x - \mu}{\sigma} \tag{17}$$

where μ is the mean of the feature, σ is the standard deviation, x means the original feature and $\hat{x}$ is after normalized one.

The MATLAB are used for the data preprocessing. The neural networks and training algorithms are implemented in Python 3.6 and Keras with Tensorflow as the back end, and the model training process is carried out on Tesla K40 GPUs. The experiments are performed on the computer of the 64-bit windows Sever 2012 R2 standard system, the CPU is the dual-core Intel CORE E5, the clock frequency of the processor is 2.1 GHz, and running memory is 128 G.

Table 3
The structure of HD-MFM.

Block	Network	Detail
Spectrogram-CNN Block	CNN	conv1: 100 3 × 3 kernels, 2 × 2 stride MaxPooling2D: 2 × 2, 1 × 1 stride conv2: 100 3 × 3 kernels, 2 × 2 stride MaxPooling2D: 2 × 2, 1 × 1 stride conv3: 100 3 × 3 kernels, 2 × 2 stride MaxPooling2D: 2 × 2, 1 × 1 stride conv4: 100 3 × 3 kernels, 2 × 2 stride MaxPooling2D: 2 × 2, 1 × 1 stride
	Flattern	Reshape into 1D vector
	DNN	Fully-connected with 2048 Hidden units Fully-connected with 1024 Hidden units
HSF-DNN Block	DNN	Fully-connected with 1024 Hidden units Fully-connected with 1024 Hidden units
MFCC-LSTM Block	RNN	LSTM with 512 Hidden units LSTM with 256 Hidden units

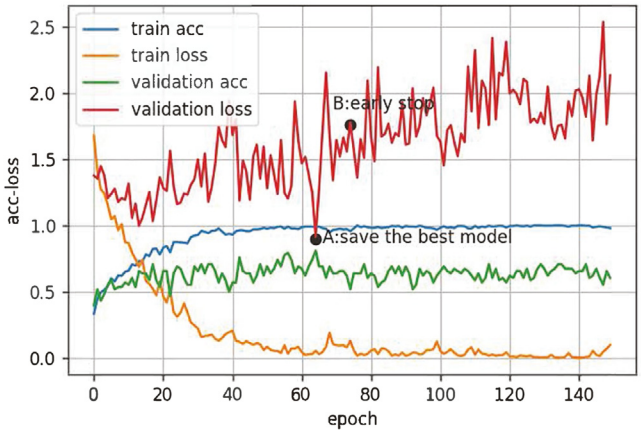


Fig. 8. The loss and accuracy of train set and validation set during the first fold training of EMO-DB in the MFCC-LSTM block. The point A means the epoch of saved best model, and the point B means the epoch of model stopping training.

The parameters of the neural networks are initialized from a Gaussian distribution with zero mean and $\sqrt{2/n_i}$ standard deviation, where n_i is the number of inputs to the layer. The batch size sets 256 and 20 for IEMOCAP and EMO-DB respectively. The number of epoch is set 150, and early stop strategy is also used in training in order to prevent over-fitting. When the loss of validation set is not decreasing within 10 epochs, network training would be stopped. Finally, the lowest loss model on validation set would be selected to test. RMSProp is used for optimizing the parameters in the DNN block, and the Adam is used both in the CNN and RNN blocks. The base learning rate is set to $\alpha = 1 * 10^{-2}$, and the decay rate is set as 0.1. The first decay of learning rate occurs after 20 epochs, when the current highest validation accuracy is reached. Start the next decay when the next time a higher accuracy is reached. The probability of dropout abandonment is 0.5.

The experiments are conducted in a speaker-independent manner. For both IEMOCAP and EMO-DB, 10-fold cross-validation is used. For each fold, the utterances from 4 sessions with total 8 speakers are used for training data, one speaker in the remaining session is used for validation and the other speaker is used for testing. When verifying the performance of the model, we use the commonly used assumptions of machine learning algorithms. That is, the training set, validation set and testing set come from the same data distribution. The entire experimental process is carried out under this assumption. When the model is trained on the training set, the best model is found through the validation set. Finally, the best model is applied to the test set to test the performance of the model. As the number of data of EMO-DB is smaller than that of IEMOCAP, overfitting is prone to occur. Therefore, the first fold training curve of EMO-DB in the MFCC-LSTM block is taken as an example. As shown in Fig. 8, the training accuracy is continuously improved with the increase of epoch, and the training loss is continuously reduced. The validation loss reaches the lowest point at point A, and it rises to cause overfitting. Therefore, the training process of three blocks and two fusion methods are in this way to prevent the model from overfitting. Then, the model stops training at point B through the early stopping mechanism.

4.2. Experiment result

To measure the performance of systems, the weighted accuracy (WA, accuracy) and unweighted accuracy (UA, average recall over different emotion categories) [64] are reported on IEMOCAP and EMO-DB. WA is the correct number of categories across the entire test set. UA represents the average result of classification accuracy for each category, which can access the impact of data category imbalances on the overall model. WA and UA can be defined as:

$$WA = \frac{\text{number of correct utterances}}{\text{number of the whole test utterances}} \tag{18}$$

$$UA = \frac{1}{K}\sum_{i=1}^{K}\frac{\text{number of correct utterances for emotion i}}{\text{number of the whole test utterances for emotion i}} \tag{19}$$

4.2.1. Selection of time duration in MFCC-LSTM block

In this part, we conduct experiments on the length of time duration selected in MFCC-LSTM Block, and due to the best performance, the sequence length is explored in the HD-MFM. Several time durations have been analyzed for this part: 50, 60, 70, 80, 90 and 100 points. The detailed result is presented in Table 4.

Table 4 shows that the best WA of 41.84%, UA of 37.79% for IEMOCAP and WA of 56.64%, UA of 52.12% for EMO-DB are both obtained if the length of time duration is set to 60. Comparing the experimental results of these two datasets, we can see that the choice of different time duration has little effect on the final accuracy, but the overall trend shows a phenomenon of first rising and then declining, and reaching a peak at 60. Hence, in the HD-MFM, the sequence length is set to 60.

4.2.2. Selection of HSFs in HSF-DNN block

In this stage, we try to find which HSF is the most suitable on the DNN. In this experiment, four commonly used HSF datasets are discussed. GeMAPS, GeMAPS with extended(eGeMAPS), interspeech′09 and interspeech′10. As mentioned above, we transform the position features of these features into position rate and remove the feature of speech length. So to the end, the dims of features form are 62, 88, 384 and 1580 respectively.

The results are showed in Table 5 and Table 6. It can be seen that when choosing the interspeech′10 dataset as the HSF, the best WA of 57.98% and UA of 58.69% can be achieved in IEMOCAP and the best WA of 82.00% and UA of 79.22% can be obtained in EMO-DB. Furthermore, considering the classification accuracy of each emotion. Interspeech′10 also has the best performance in terms of experimental results except that it is not good for a certain emotion, but it performs best from the average point of view. Thus, Interspeech′10 was selected as the HSF used in IEMOCAP and EMO-DB.

4.2.3. Recognition accuracy of HD-MFM

Since the features of the three modules contain voice information of different dimensions, this paper attempts to complement the advantages of three dimensions features in a certain way. In this experiment, we compare two different training methods, merging and separating. The comparison of the results of the two combination methods are shown in Table 7. Obviously, when compared with the method of merging, the separating shows a certain advantage in overall performance. The WA and the UA achieved by learning deep features from the separating method are much higher than that from the merging method. Moreover, the separating method can learn the high-dimensional feature representations acquired through the above three blocks, which greatly improves the ability of emotional recognition. The ability of the merging method to recognize emotions is not even as good as HSF-DNN block does. The reason may be that in the merging method, three blocks are trained in the global model at the same time, and each block is easily affected by the training of others. Because the loss is shared globally, the convergence of the three

Table 4
Performance (%) comparison between the different Sequence length for LSTM on the IEMOCAP and EMO-DB dataset (Ang = Angry, Hap = Happy, Neu = Neutral, Sad = Sadness).

DB	IEMOCAP						EMO-DB					
Seq length	WA	UA	Ang	Hap	Neu	Sad	WA	UA	Ang	Hap	Neu	Sad
50	41.15	36.81	19.84	43.31	**48.08**	35.99	54.68	47.74	74.57	18.97	53.69	43.71
60	**41.84**	**37.79**	21.34	**47.26**	44.05	38.50	**56.64**	**52.12**	72.17	18.81	**58.26**	**59.22**
70	40.28	37.00	**22.13**	44.55	44.47	36.85	54.45	47.04	**77.53**	18.57	55.69	36.36
80	40.05	37.11	21.61	44.31	43.12	**39.39**	54.92	48.33	76.00	19.52	54.25	43.54
90	40.01	36.98	21.16	45.53	42.74	38.47	54.82	48.47	75.45	**19.69**	57.88	40.84
100	39.94	36.90	20.71	45.88	42.17	38.84	54.34	47.50	75.81	19.62	55.93	38.63
AVG	40.55	37.10	21.13	45.14	44.11	38.01	54.98	48.53	75.26	19.20	55.95	43.72

Table 5
Performance (%) comparison between the different HSF datasets for DNN on the EMO-DB dataset (Ang = Angry, Hap = Happy, Neu = Neutral, Sad = Sadness).

EMO-DB	WA	UA	Ang	Hap	Neu	Sad
Interspeech'09	79.92	78.59	87.56	56.36	87.57	**82.88**
Interspeech'10	**82.00**	**79.22**	**94.62**	**57.34**	**87.77**	77.16
GeMAPs	73.95	72.45	82.61	45.18	82.63	79.38
eGeMAPs	76.18	72.02	90.96	39.20	79.97	77.96

Table 6
Performance (%) comparison between the different HSF datasets for DNN on the IEMOCAP dataset (Ang = Angry, Hap = Happy, Neu = Neutral, Sad = Sadness).

IEMOCAP	WA	UA	Ang	Hap	Neu	Sad
Interspeech'09	54.51	52.83	57.32	54.08	45.74	54.16
Interspeech'10	**57.98**	**58.69**	**61.49**	**61.20**	44.32	**67.76**
GeMAPs	53.54	53.10	51.38	50.80	45.45	64.77
eGeMAPs	54.81	54.23	51.29	56.87	**46.79**	61.96

Table 7
Performance (%) of separating and merging methods in HD-MFM for IEMOCAP and EMODB (Ang = Angry, Hap = Happy, Neu = Neutral, Sad = Sadness).

DB	Method	WA	UA	ang	hap	neu	sad
IEMOCAP	Separating	**72.02**	**73.42**	**78.30**	**68.51**	**64.70**	**82.17**
	Merging	56.30	56.29	69.56	46.11	55.69	53.81
EMO-DB	Separating	**91.25**	**90.61**	**88.49**	**86.06**	**92.88**	**95.00**
	Merging	75.53	72.39	93.16	42.41	77.61	76.38

Table 8
Performance (%) comparison between the different blocks and two fusion methods on the EMO-DB dataset and IEMOCAP dataset with WA and UA.

Method	EMO-DB		IEMOCAP	
	WA (%)	UA (%)	WA (%)	UA (%)
Spectrogram-CNN Block	50.01	61.75	48.67	45.54
HSF-DNN Block	82.00	79.22	57.98	58.69
MFCC-LSTM Block	56.64	52.12	41.84	37.79
Merging	75.53	72.39	56.30	56.29
Separating	**91.25**	**90.61**	**72.02**	**73.42**

blocks may not be as advantageous as the independent training of the modules. All in all, the merging method does not make good use of high-dimensional features. In contrast, the separating method can ensure that the independent features of each block are not affected by other blocks. The features extracted by the three trained independent models retain the speech information of different representations. Then the three feature vectors are concatenated to get the final complementary feature for SER training. In a nutshell, separating can fully integrate the features of the three blocks to further enhance the capabilities of SER.

The independent results of the three blocks and two fusion methods on EMO-DB and IEMOCAP are shown in the Table 8 and the confusion matrices of the first fold are shown in Figs. 10 and 11. The deep learning models trained by different representations of speech have obvious differences in the effect of SER. On the EMO-DB database, HSF-DNN has obvious advantages compared to the Spectrogram-CNN and MFCC-LSTM. This shows that the features extracted by HSF-DNN can more effectively express speech emotional information. At the same time, Spectrogram-CNN and MFCC-LSTM can also get acceptable results, although not as good as HSF-DNN. On the IEMOCAP database, the results of Spectrogram-CNN show that the four emotion classifications are relatively more uniform. The result of MFCC-LSTM has a great improvement on Neu emotion. HSF-DNN has obvious advantages in the other three emotions. The experimental results show that different forms of speech data representations have different sensitivities to different emotions. Finally, from Figs. 10(e) and 11(e), the proposed HD-MFM can make full use of the complementarity between three blocks, thus obtaining a better set of feature representations, which significantly improves the accuracy of SER.

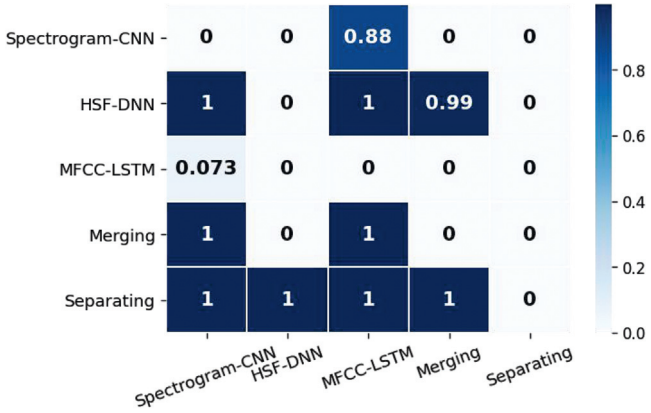


Fig. 9. The probability metrics from Bayesian signed rank test on three proposed blocks and two fusion methods.

To further validate the performance of proposed separating method, Bayesian analysis with the value of $\rho = 1\%$ [65] is necessary. The Bayesian analysis is used to compare the WA and UA classification results on EMO-DB and IEMOCAP among different methods. In the Bayesian matrix, each value represents the probability that the row method is better than the column method. By the way, the greater the probability, the higher the value, and the darker the color. As is shown in Fig. 9, the Bayesian probabilities are all as high as 1, when the WA and UA of separating on both datasets. The separating method has obvious advantages in SER. Compared with the merging method and the three blocks, the separating method can fully characterize the speech information to complete the SER.

The recognition rate on EMO-DB and IEMOCAP obtained from the proposed method was compared with the previous research works. All methods follow the same experimental strategy,

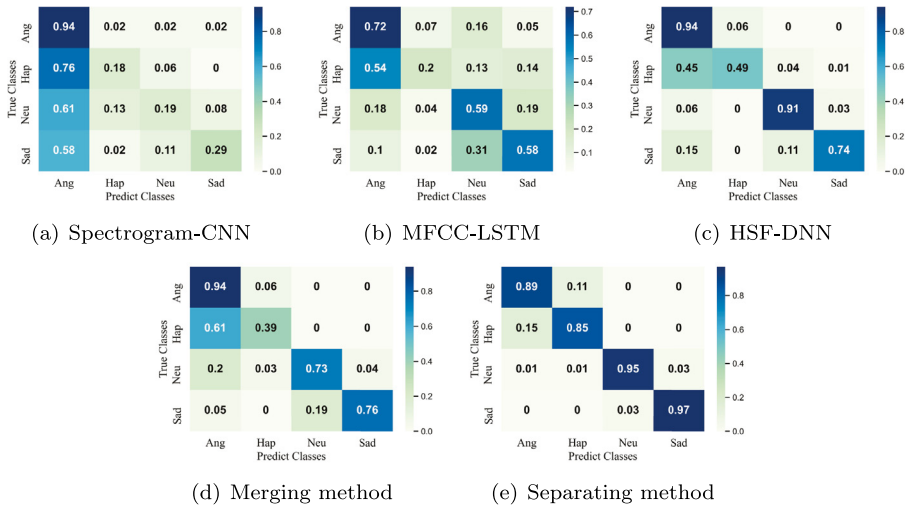


Fig. 10. Confusion matrices of three blocks and two combining method in EMO-DB (Ang = Angry, Hap = Happy, Neu = Neutral, Sad = Sadness).

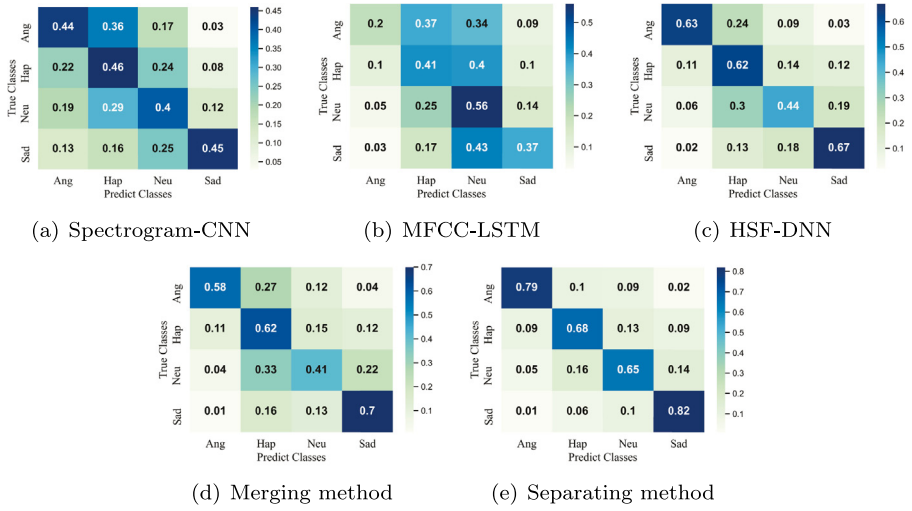


Fig. 11. Confusion matrices of three blocks and two combining method in IEMOCAP (Ang = Angry, Hap = Happy, Neu = Neutral, Sad = Sadness).

which is 10 folds LOSO cross-validation based on the speaker-independent, which is the same setup as ours. The results have been tabulated in Tables 10 and 9. Since the detail experimental parameters of other methods are not given in their papers, we only provide the overall trainable parameters of our method and the trainable parameters of [69] for the actual number of trainable parameters. The proposed HD-MFM shows great advantages on the EMODB dataset. Although the WA performance on IEMOCAP is smaller than other methods, the UA maintains a higher value. Because WA represents the average accuracy rate of all data, and UA represents average recall over different emotion categories, and HD-MFM achieves a balance between the effects of different emotional categories. For the methods [67], we obtained the emotion recognition results using only speech data. Further from Table 10, it can also be concluded that the proposed HD-MFM has obtained a much higher recognition accuracy in SER by comprehensively considering rich speech feature representations. Although we use a lot of features and neural networks with more trainable parameters, HD-MFM shows the excellent results, which is acceptable. In general, the proposed HD-MFM considers the complementarity of different types of speech features, and utilizes different neural networks to extract different types of speech features to further improve the ability of the model to recognize speech emotion.

5. Conclusion and future work

The emotional information contained in different forms of speech features is different and complementary. Building different models to obtain different information in these features may be beneficial for improving the accuracy of emotional recognition. Thus, in this paper, we propose a framework, which combines three different blocks to learn the speech features of three dimensions, namely HD-MFM. Three dimensions features of speech are combined by separating strategy to obtain the higher performance. HD-MFM utilizes the complementarity of speech features from three different dimensions, image, temporal and statistical speech feature dimension, and combines the respective characteristics of three different neural networks, CNN, DNN and LSTM to obtain a high-dimensional speech feature. The performance of the HD-MFM is tested on IEMOCAP and EMODB databases. HD-MFM obtains **72.02%** WA and **73.42%** UA in IEMOCAP, and gets **91.25%** WA and **90.61%** UA in EMODB. The experiments show that the HD-MFM can learn distinguishing features and provide more accurate predictions compared with those three original blocks. We also compare with other well-established feature representations and methods, which shows the HD-MFM also has significant advantages.

Table 9
Comparison of the used features and neural networks with previous works.

Method	Speech features	Neural networks
Gangamohan P. et al.(2015) [66]	Fundamental frequency (F0), strength of excitation (SoE) energy of excitation (EoE)	–
Z. Peng et al.(2021) [67]	MFCC, X-vector	CNN
W. Fan et al.(2020) [42]	Speech spectrogram	CNN
H. Li et al.(2020) [43]	Log-mel filterbank energies (Log-MFBs), pitch,energy	CNN, DNN
F. Daneshfar et al.(2020) [18]	MFCC, perceptual linear prediction cepstral coefficients (PLPC), perceptual minimum-variance distortionless response cepstral coefficients (PMVDR), pitch and their first and second order derivatives	–
S. Zhong et al.(2020) [68]	MFCC, openSMILE features	CNN, LSTM, DNN
S. Lee et al.(2020) [34]	Speech spectrogram	CNN
D. Li et al.(2021) [69]	MFCC, spectral roll-off point, spectral flux, spectral centroid, spectral entropy, spectral spread, zero-crossing rate, fundamental frequency, energy, energy entropy and their first-order difference	CNN, LSTM (total 2,340,740 trainable parameters)
Proposed	Speech spectrogram, MFCC, interspeech'10	CNN, LSTM, DNN (total 12,417,756 trainable parameters)

Table 10
Comparison of the proposed results with previous work results. (accuracy (%)).

Method	Accuracy			
	EMODB		IEMOCAP	
	WA (%)	UA (%)	WA (%)	UA (%)
Gangamohan P. et al.(2015) [66]	75.22	–	–	–
Z. Peng et al.(2021) [67]	–	–	66.6	68.4
W. Fan et al.(2020) [42]	–	–	73.02	65.86
H. Li et al.(2020) [43]	–	–	58.62	59.91
F. Daneshfar et al.(2020) [18]	82.82	–	74.80	–
S. Zhong et al.(2020) [68]	85.76	86.12	**74.98**	68.83
S. Lee et al.(2020) [34]	88.43	86.04	66.47	67.12
D. Li et al.(2021) [69]	85.95	82.06	61.20	54.99
Proposed	**91.25**	**90.61**	72.02	**73.42**

Despite the good experimental results of HD-MFM, there has still an aspect need to be improved. That is, the imbalance problem of the SER. In the real situation, the emotional proportion in the dialogue between people is different, such as more neutral emotions and less angry emotions. Therefore, many speech data sets are recorded by copying this ratio, and so is IEMOCAP. This imbalance factor makes the recognition rate of a particular emotion low. This experiment shows that the recognition rate of happiness is low. Therefore, if the recognition rate of happy emotion in this experiment is improved, the overall accuracy of emotion will be significantly improved. Hence in the future, we plan to explore more effective ways of combining information from the specific emotional dimension. Besides, we will further consider the feature selection of different types of speech features, and utilize feature selection algorithms to improve the ability of speech emotion recognition.

CRediT authorship contribution statement

Xinlei Xu: Writing – original draft, Writing – review & editing, Experiment. **Dongdong Li:** Investigation, Methodology. **Yijun Zhou:** Validation. **Zhe Wang:** Project administration, Proofreading.

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

The two experimental datasets used in this paper are publicly available online.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No. 62276098 and No. 62076094, Shanghai Science and Technology Program, China "Federated based cross-domain and cross-task incremental learning" under Grant No. 21511100800, Shanghai Science and Technology Program, China "Distributed and generative few-shot algorithm and theory research" under Grant No. 20511100600, Chinese Defense Program of Science and Technology under Grant No. 2021-JCJQ-JJ-0041, China Aerospace Science and Technology Corporation Industry-University-Research Cooperation Foundation of the Eighth Research Institute under Grant No. SAST2021-007.

References

[1] B. Huebner, S. Dwyer, M. Hauser, The role of emotion in moral psychology, Trends Cogn. Sci. 13 (1) (2009) 1–6.
[2] S.L. Koole, The psychology of emotion regulation: An integrative review, Cogn. Emot. 23 (1) (2009) 4–41.
[3] P.M. Niedenthal, F. Ric, Psychology of Emotion, Psychology Press, 2017.
[4] S. Ramakrishnan, I.M.M.E. Emary, Speech emotion recognition approaches in human computer interaction, Telecommun. Syst. 52 (3) (2013) 1467–1478.
[5] M.S. Fahad, A. Ranjan, J. Yadav, A. Deepak, A survey of speech emotion recognition in natural environment, Digit. Signal Process. (2020) 102951.
[6] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, X. Xu, Spatiotemporal and frequential cascaded attention networks for speech emotion recognition, Neurocomputing 448 (2021) 238–248.

[7] K.L. de Ipiña, J.B. Alonso, N. Barroso, M. Faúndez-Zanuy, M. Ecay, J. Solé-Casals, C.M. Travieso, A. Estanga, A. Ezeiza, New approaches for Alzheimer's disease diagnosis based on automatic spontaneous speech analysis and emotional temperature, in: IWAAL, Vitoria-Gasteiz, Spain, 2012, pp. 407–414.

[8] R. Khokale, A.R. Panat, Y.H. Gulhane, Analysis of affective speech for fatigue detection, in: Proc. ICWET '10 Int. Conf. & Work. Emerg. Trends Technol. Mumbai, Maharashtra, India, 2010, pp. 237–240.

[9] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, in: IEEE Int. Conf. Acoust. Speech Signal Process., 2013, pp. 3687–3691.

[10] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, R. Prasad, Ensemble of SVM trees for multimodal emotion recognition, in: APSIPA 2012, Hollywood, CA, USA, December 3-6, 2012, 2012, pp. 1–4.

[11] C.-N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artif. Intell. Rev. 43 (2) (2015) 155–177, http://dx.doi.org/10.1007/s10462-012-9368-5.

[12] A.B. Kandali, A. Routray, T.K. Basu, Emotion recognition from assamese speeches using MFCC features and GMM classifier, in: TENCON 2008 - 2008 IEEE Reg. 10 Conf., 2008, pp. 1–5, http://dx.doi.org/10.1109/TENCON.2008.4766487.

[13] M. Vondra, R. Vích, Evaluation of speech emotion classification based on GMM and data fusion, in: Cross-Modal Anal. Speech, Gestures, Gaze Facial Expressions, COST Action 2102 Int. Conf. Prague, Czech Republic, Oct. 15-18, 2008, Revis. Sel. Invit. Pap., 2008, pp. 98–105, http://dx.doi.org/10.1007/978-3-642-03320-9_10.

[14] J. Yadav, K.S. Rao, Neural network and GMM based feature mappings for consonant-vowel recognition in emotional environment, Int. J. Speech Technol. 21 (3) (2018) 421–433, http://dx.doi.org/10.1007/s10772-017-9478-1.

[15] D. Le, E.M. Provost, Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks, in: 2013 IEEE Work. Autom. Speech Recognit. Underst., 2013, pp. 216–221, http://dx.doi.org/10.1109/ASRU.2013.6707732.

[16] Y. Ijima, M. Tachibana, T. Nose, T. Kobayashi, Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM, in: Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. ICASSP 2009, 19-24 April 2009, Taipei, Taiwan, 2009, pp. 4157–4160, http://dx.doi.org/10.1109/ICASSP.2009.4960544.

[17] J. Lorenzo-Trueba, R. Barra-Chicote, R.S. Segundo, J. Ferreiros, J. Yamagishi, J.M. Montero, Emotion transplantation through adaptation in HMM-based speech synthesis, Comput. Speech Lang. 34 (1) (2015) 292–307, http://dx.doi.org/10.1016/j.csl.2015.03.008.

[18] F. Daneshfar, S.J. Kabudian, Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm, Multimedia Tools Appl. 79 (1–2) (2020) 1261–1289.

[19] N. Yang, J. Yuan, Y. Zhou, I. Demirkol, Z. Duan, W.B. Heinzelman, M. Sturge-Apple, Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification, Int. J. Speech Technol. 20 (1) (2017) 27–41, http://dx.doi.org/10.1007/s10772-016-9364-2.

[20] H. Aouani, Y.B. Ayed, Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder, in: 4th Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2018, Sousse, Tunis. March 21-24, 2018, 2018, pp. 1–5, http://dx.doi.org/10.1109/ATSIP.2018.8364518.

[21] S. Demircan, H. Kahramanli, Emotion recognition from assamese speeches using MFCC features and GMM classifier, in: J. Adv. Comput. Networks, Vol. 2, 2014, pp. 28–30, http://dx.doi.org/10.7763/JACN.2014.V2.76.

[22] M.T. Shami, M.S. Kamel, Segment-based approach to the recognition of emotions in speech, in: Proc. 2005 IEEE Int. Conf. Multimed. Expo, ICME 2005, July 6-9, 2005, Amsterdam, Netherlands, 2005, pp. 366–369, http://dx.doi.org/10.1109/ICME.2005.1521436.

[23] Z. Liu, M. Wu, W. Cao, J.-W. Mao, J.-P. Xu, G. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280, http://dx.doi.org/10.1016/j.neucom.2017.07.050.

[24] G.E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554, http://dx.doi.org/10.1162/neco.2006.18.7.1527.

[25] O.A. Alzubi, J.A. Alzubi, M. Alweshah, I. Qiqieh, S. Al-Shami, M. Ramachandran, An optimal pruning algorithm of classifier ensembles: dynamic programming approach, Neural Comput. Appl. 32 (20) (2020) 16091–16107.

[26] A.A. Movassagh, J.A. Alzubi, M. Gheisari, M. Rahimi, S. Mohan, A.A. Abbasi, N. Nabipour, Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model, J. Ambient Intell. Humaniz. Comput. (2021) 1–9.

[27] Y. Bengio, A.C. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828, http://dx.doi.org/10.1109/TPAMI.2013.50.

[28] L. Chen, X. Mao, H. Yan, Text-independent phoneme segmentation combining EGG and speech data, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (6) (2016) 1029–1037, http://dx.doi.org/10.1109/TASLP.2016.2533865.

[29] K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, in: INTERSPEECH 2014, 15th Annu. Conf. Int. Speech Commun. Assoc. Singapore, Sept. 14-18, 2014, 2014, pp. 223–227.

[30] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using CNN, in: Proc. ACM Int. Conf. Multimedia, MM '14, Orlando, FL, USA, Novemb. 03 - 07, 2014, 2014, pp. 801–804, http://dx.doi.org/10.1145/2647868.2654984.

[31] J. Zhao, X. Mao, L. Chen, Learning deep features to recognise speech emotion using merged deep CNN, IET Signal Process. 12 (6) (2018) 713–721, http://dx.doi.org/10.1049/iet-spr.2017.0320.

[32] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomed. Signal Proc. Control 47 (2019) 312–323, http://dx.doi.org/10.1016/j.bspc.2018.08.035.

[33] S. Kwon, et al., MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach, Expert Syst. Appl. 167 (2021) 114177.

[34] S. Lee, D.K. Han, H. Ko, Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition, Sensors 20 (22) (2020) 6688.

[35] U. Kumaran, S.R. Rammohan, S.M. Nagarajan, A. Prathik, Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN, Int. J. Speech Technol. 24 (2) (2021) 303–314.

[36] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[37] F.A. Gers, J. Schmidhuber, F.A. Cummins, Learning to forget: Continual prediction with LSTM, Neural Comput. 12 (10) (2000) 2451–2471, http://dx.doi.org/10.1162/089976600300015015.

[38] J.A. Alzubi, R. Jain, A. Kathuria, A. Khandelwal, A. Saxena, A. Singh, Paraphrase identification using collaborative adversarial networks, J. Intell. Fuzzy Syst. 39 (1) (2020) 1021–1032.

[39] J.A. Alzubi, R. Jain, P. Nagrath, S. Satapathy, S. Taneja, P. Gupta, Deep image captioning using an ensemble of CNN and LSTM based deep neural networks, J. Intell. Fuzzy Syst. (Preprint) (2021) 1–9.

[40] K. Huang, C. Wu, T. Yang, M. Su, J. Chou, Speech emotion recognition using autoencoder bottleneck features and LSTM, in: Int. Conf. Orange Technol., 2016, pp. 1–4, http://dx.doi.org/10.1109/ICOT.2016.8278965.

[41] Z. Peng, J. Dang, M. Unoki, M. Akagi, Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech, Neural Netw. 140 (2021) 261–273.

[42] W. Fan, X. Xu, X. Xing, D. Huang, Adaptive domain-aware representation learning for speech emotion recognition, in: INTERSPEECH, 2020, pp. 4089–4093.

[43] H. Li, M. Tu, J. Huang, S. Narayanan, P. Georgiou, Speaker-invariant affective representation learning via adversarial training, in: ICASSP 2020-2020 IEEE Int. Conf. Acoust. Speech Signal Process., IEEE, 2020, pp. 7144–7148.

[44] D. Tang, J. Zeng, M. Li, An end-to-end deep learning framework for speech emotion recognition of atypical individuals, in: Interspeech, 2018, pp. 162–166, http://dx.doi.org/10.21437/Interspeech.2018-2581.

[45] D. Luo, Y. Zou, D. Huang, Investigation on joint representation learning for robust feature extraction in speech emotion recognition, in: Interspeech, 2018, pp. 152–156, http://dx.doi.org/10.21437/Interspeech.2018-1832.

[46] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, C. Li, Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition, in: Interspeech, 2018, pp. 272–276, http://dx.doi.org/10.21437/Interspeech.2018-1477.

[47] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, Convolutional neural networks for speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (10) (2014) 1533–1545, http://dx.doi.org/10.1109/TASLP.2014.2339736.

[48] J. Kaur, A. Kumar, Speech emotion recognition using CNN, k-NN, MLP and random forest, in: Computer Networks and Inventive Communication Technologies, Springer, 2021, pp. 499–509.

[49] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, B. Schuller, Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition, IEEE Access 7 (2019) 97515–97525.

[50] L. Xu, C.-s. Choy, Y.-W. Li, Deep sparse rectifier neural networks for speech denoising, in: IEEE Int. Work. Acoust. Signal Enhanc. IWAENC 2016, Xi'an, China, Sept. 13-16, 2016, 2016, pp. 1–5, http://dx.doi.org/10.1109/IWAENC.2016.7602891.

[51] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. Andr??, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, K.P. Truong, The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202, http://dx.doi.org/10.1109/TAFFC.2015.2457417.

[52] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, in: IEEE Int. Conf. Acoust. Speech Signal Process., 2013, pp. 3687–3691, http://dx.doi.org/10.1109/ICASSP.2013.6638346.
[53] Y. Sun, G. Wen, J. Wang, Weighted spectral features based on local Hu moments for speech emotion recognition, Biomed. Signal Proc. Control 18 (2015) 80–90, http://dx.doi.org/10.1016/j.bspc.2014.10.008.
[54] S.A.A. Yusuf, R. Hidayat, MFCC feature extraction and KNN classification in ECG signals, in: 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), IEEE, 2019, pp. 1–5.
[55] N.A. Zaidan, M.S. Salam, MFCC global features selection in improving speech emotion recognition rate, in: Advances in Machine Learning and Signal Processing, Springer, 2016, pp. 141–153.
[56] H. Isyanto, A.S. Arifin, M. Suryanegara, Voice biometrics for Indonesian language users using algorithm of deep learning CNN residual and hybrid of DWT-MFCC extraction features, Int. J. Adv. Comput. Sci. Appl. 13 (5) (2022).
[57] B.S. Soares, J.S. Luz, V.F. de Macêdo, R.R.V. e Silva, F.H.D. de Araújo, D.M.V. Magalhães, MFCC-based descriptor for bee queen presence detection, Expert Syst. Appl. 201 (2022) 117104.
[58] P.-W. Hsiao, C.-P. Chen, Effective attention mechanism in dynamic models for speech emotion recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 2526–2530.
[59] H.M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Netw. 92 (2017) 60–68.
[60] Y. Kim, E.M. Provost, Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions, in: IEEE Int. Conf. Acoust. Speech Signal Process., 2013, pp. 3677–3681, http://dx.doi.org/10.1109/ICASSP.2013.6638344.
[61] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359, http://dx.doi.org/10.1007/s10579-008-9076-6.
[62] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech, in: INTERSPEECH 2005 - Eurospeech, 9th Eur. Conf. Speech Commun. Technol. Lisbon, Port. Sept. 4-8, 2005, 2005, pp. 1517–1520.
[63] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, in: ACM Int. Conf. Multimed., 2013, pp. 835–838.
[64] B.W. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2009 emotion challenge, in: INTERSPEECH 2009, 10th Annu. Conf. Int. Speech Commun. Assoc. Bright. United Kingdom, Sept. 6-10, 2009, 2009, pp. 312–315.
[65] A. Benavoli, G. Corani, J. Demšar, M. Zaffalon, Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis, J. Mach. Learn. Res. 18 (1) (2017) 2653–2688.
[66] S.R. Kadiri, P. Gangamohan, S.V. Gangashetty, B. Yegnanarayana, Analysis of excitation source features of speech for emotion recognition, in: Sixt. Annu. Conf. Int. Speech Commun. Assoc., 2015.
[67] Z. Peng, Y. Lu, S. Pan, Y. Liu, Efficient speech emotion recognition using multi-scale CNN and attention, in: ICASSP 2021-2021 IEEE Int. Conf. Acoust. Speech Signal Process., IEEE, 2021, pp. 3020–3024.
[68] S. Zhong, B. Yu, H. Zhang, Exploration of an independent training framework for speech emotion recognition, IEEE Access 8 (2020) 222533–222543.
[69] D. Li, J. Liu, Z. Yang, L. Sun, Z. Wang, Speech emotion recognition using recurrent neural networks with directional self-attention, Expert Syst. Appl. 173 (2021) 114683.