Speech Communication 127 (2021) 73–81

Contents lists available at ScienceDirect

Speech Communication

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

Learning deep multimodal affective features for spontaneous speech emotion recognition

Shiqing Zhang, Xin Tao, Yuelong Chuang, Xiaoming Zhao *

Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, PR China

ARTICLE INFO

Keywords:
Speech emotion recognition
Convolutional neural networks
Deep multimodal feature learning
Temporal-spatial
Score-level fusion

ABSTRACT

Recently, spontaneous speech emotion recognition has become an active and challenging research subject. This paper proposes a new method of spontaneous speech emotion recognition by using deep multimodal audio feature learning based on multiple deep convolutional neural networks (multi-CNNs). The proposed method initially generates three different audio inputs for multi-CNNs so as to learn deep multimodal segment-level features from the original 1D audio signal in three aspects: 1) a 1D CNN for 1D raw waveform modeling, 2) a 2D CNN for 2D time-frequency Mel-spectrogram modeling, and 3) a 3D CNN for temporal-spatial dynamic modeling. Then, an average-pooling is performed on the obtained segment-level classification results from 1D, 2D, and 3D CNN networks, to produce utterance-level classification results. Finally, a score-level fusion strategy is adopted as a multi-CNN fusion method to integrate different utterance-level classification results for final emotion classification. The learned deep multimodal audio features are shown to be complementary to each other so that they are combined in a multi-CNN fusion network to achieve significantly improved emotion classification performance. Experiments are conducted on two challenging spontaneous emotional speech datasets, *i.e.*, the AFEW5.0 and BAUM-1 s databases, demonstrating the promising performance of our proposed method.

1. Introduction

In recent years, spontaneous speech emotion recognition (SER) has become an active and challenging research subject in pattern recognition, speech signal processing, artificial intelligence, and so on. This is because spontaneous SER has important applications to human-computer interaction (Li et al., 2018; Zhang et al., 2013). In particular, these SER systems aim to provide affective interaction modes with computers by direct speech interaction rather than traditional input devices, thereby giving rise to smart affective services for spoken call center, healthcare, surveillance, and affective computing.

In SER community, a variety of previous works (Akçay and Oğuz, 2020; Anagnostopoulos et al., 2015; El Ayadi et al., 2011; Liu et al., 2018; Schuller, 2018; Wang et al., 2020) concentrate on acted SER tasks on the basis of collected data related to acted emotion expression. The main reason is that acted emotions are easily portrayed in the laboratory controlled environment, and usually produce good SER performance. However, acted emotions are often exaggerated so that they cannot effectively represent the characteristics of speech emotion expression in real-world sceneries. Therefore, identifying spontaneous emotions in the wild is more difficult and challenging compared with conventional acted emotions.

Speech feature extraction, which is a crucial step in a basic spontaneous SER system, aims to derive effective feature representations related to speech emotion expression. The well-known affective speech features (Demircan and Kahramanli, 2017; Gharavian et al., 2012; Song, 2019; Zhang et al., 2018a; Zhao and Zhang, 2015; Zixing et al., 2015) are low-level descriptors (LLDs). The early-used typical LLDs contain prosody (pitch, intensity) features, voice quality (formants) features, spectral features such as Mel-frequency cepstral coefficients (MFCCs), linear predictor coefficients (LPC), and linear predictor cepstral coefficients (LPCC). Recently, various extensive feature sets on the basis of LLDs, including INTERSPEECH-2010 (Kayaoglu and Eroglu Erdem, 2015), ComParE (Schuller et al., 2013), AVEC-2013 (Valstar et al., 2013), and GeMAPS (Eyben et al., 2016), have been also developed for SER. Nevertheless, all these extracted LLDs and its variants belong to low-level hand-designed features. Due to the gap between low-level hand-designed features and subjective emotions, these

* Corresponding author.
E-mail address: tzxyzxm@163.com (X. Zhao).

https://doi.org/10.1016/j.specom.2020.12.009
Received 23 May 2020; Received in revised form 2 November 2020; Accepted 21 December 2020
Available online 26 December 2020

low-level hand-designed features are not effective enough to represent emotional characteristics of speech (Akçay and Oğuz, 2020; Anagnostopoulos et al., 2015; El Ayadi et al., 2011; Liu et al., 2018; Schuller, 2018; Wang et al., 2020). Accordingly, it is desired to develop advanced feature learning approaches to automatically achieve high-level affective feature representations that characterize speakers' emotions.

To address the above-mentioned issues, the recently-emerged deep learning techniques (Hinton and Salakhutdinov, 2006; LeCun et al., 2015), which have gained extensive attentions in SER community, may offer possible solutions. Due to the used deep layers of architectures, deep learning methods usually have certain advantages over traditional methods, including their ability to automatically detect the complicated structures and features without manual feature extraction. So far, various deep learning techniques, such as deep neural networks (DNNs) (Hinton and Salakhutdinov, 2006), deep convolutional neural networks (CNNs) (Krizhevsky et al., 2012), long short-term memory based recurrent neural networks (LSTM-RNNs) (Graves, 2012), have been used for high-level feature learning tasks for SER. On the basis of feed-forward structures, DNNs contain one or more underlying hidden layers between inputs and outputs. The feed-forward architectures make DNNs present promising performance for SER. In detail, in (Han et al., 2014), MFCCs are fed into a DNN for learning high-level features, and then an extreme learning machine (ELM) is used for speech emotion classification. In (Wang and Tashev, 2017), a DNN is used to encode all frames in an utterance into a fixed-length vector by pooling the activations of the last hidden layer over time. A kernel ELM is further trained on the encoded vectors for utterance-level emotion classification. Due to the used hand-designed features as inputs of DNNs, DNNs cannot effectively obtain discriminative features for SER.

CNNs comprise of multi-level convolutional and pooling layers, so that they are capable of capturing mid-level feature representations from input data. Benefited from the obtained great success of CNNs in computer vision tasks (Krizhevsky et al., 2012), a 2D time-frequency representation derived from an audio spectrogram is usually fed into CNNs for SER. In particular, in (Mao et al., 2014) the authors adopt audio spectrograms as inputs of a hybrid deep model, which combines a sparse auto-encoder with a 1-layer CNN, to learn salient features for SER. In (Badshah et al., 2019), segment-level spectrograms are fed into a CNN consisting of five convolutional layers and three pooling layers, to capture discriminative features for SER. In (Zhang et al., 2018c), an image-like spectrogram is developed as inputs of a deep CNN like AlexNet (Krizhevsky et al., 2012) to extract high-level segment-level feature representations for SER. In recent years, combining CNNs with LSTM-RNNs (*i.e.*, CNN+LSTM/RNNs) has become a new research trend in SER community. In (Zhao et al., 2019b, 2018), by using segment-level spectrograms, the authors integrate an attention-based bidirectional LSTM with a spatial CNN with a fully convolutional networks (FCN) like structure for deep spectrum features extraction on SER tasks. In (Zhang et al., 2019a), a multiscale deep convolutional LSTM framework based on segment-level spectrograms is presented for SER. In (Zhao et al., 2019a), the authors provide compact convolutional RNNs via binarization for SER by means of quantizing the weights of neural networks from the original full-precised values into binary values.

Note that the above-mentioned 2D CNNs-based methods, such as CNNs and CNN+LSTM/RNNs, have the ability to capture energy modulation patterns across time and frequency from the extracted 2D time-frequency spectrograms of audio signals, and hence achieve good performance on SER tasks. However, such 2D CNNs-based methods employ 2D time-frequency spectrograms as inputs of CNNs to learn feature representations. Consequently, they fail to capture the changes in 2D time-frequency representations of consecutive frames in an utterance, thereby failing to obtain discriminative enough features for SER. Although LSTM-RNNs can be used for temporal modeling of audio signals, they tend to overemphasize the temporal information.

To tackle this issue, recently-developed 3D CNNs (Dong et al., 2020; Tran et al., 2015), originally used for video processing, may provide possible solutions, since 3D CNNs are able to simultaneously learn temporal and spatial feature representations by means of 3D convolution and pooling operations. Motivated by 3D motions of videos, we will generate appropriate 3D signals from the original 1D audio signals as inputs of 3D CNNs. Extracting such video-like 3D signals aims to emphasize different spectral characteristics from neighboring regions in an utterance in the temporal and spatial dimension.

Additionally, in recent years different 1D CNN models have also been leveraged for feature learning on SER tasks. In (Fayek et al., 2017), the authors focus on investigating the performance of multiple 1D CNN structures, which contain one or two convolution layers, for learning feature representations from the original 1D raw audio waveforms on SER tasks. Nevertheless, these used 1D CNN models with one or two convolution layers are relatively shallow, so that their learned 1D CNN features may not be discriminative enough for SER. To address this issue, the recently-developed deep sample-level 1D CNNs (Kim et al., 2018; Lee et al., 2018), in which the filter size in the bottom layer to span several samples long, may provide an solution. To date, sample-level 1D CNNs have been successfully employed to learn feature representations based on the original 1D raw audio waveforms on music classification and generation tasks. Motivated by the VGG networks (Simonyan and Zisserman, 2015) in image classification that is built with deep stack of small convolutional layers, sample-level CNNs adopt very small filters in time for all convolutional layers, and show comparable performance obtained by 2D Mel-spectrograms in music classification and generation (Kim et al., 2018; Lee et al., 2018).

It is noted that the learned feature representations from 1D, and 3D CNNs may capture quite different acoustic characteristics in comparison with time-frequency representations based 2D CNNs. In particular, the raw 1D audio waveforms as inputs of 1D CNNs, are used to address log-scale amplitude compression and phase-invariance (Kim et al., 2018; Lee et al., 2018). The extracted 2D Mel-spectrogram segments as inputs of 2D CNNs, are used to capture energy modulation patterns across time and frequency (Zhang et al., 2018c). The extracted 3D video-like Mel--spectrogram segments as inputs of 3D CNNs, aim to emphasize different spectral characteristics from neighboring regions in an utterance in the temporal and spatial dimension. This demonstrate that the learned deep multimodal features from 1D, 2D, and 3D CNNs, may be complementary to each other, so that they are integrated in a multi-CNN fusion network to further improve speech emotion classification performance.

Motivated by the existing complementarity among the learned deep multimodal features from 1D, 2D, and 3D CNN networks, we propose a new spontaneous SER method, which aims to learn deep multimodal audio features with a multi-CNN fusion network to make potential performance improvements on spontaneous SER tasks. Fig. 1 shows the overview architecture of our proposed method. In particular, we propose to generate three appropriate audio inputs corresponding to three different CNN architectures, so as to learn deep multimodal features from the original 1D audio signals in three aspects: 1) a 1D CNN for 1D raw waveform modeling, 2) a 2D CNN for 2D time-frequency Mel-spectrogram modeling, and 3) a 3D CNN for temporal-spatial dynamic modeling.

The main contributions of this paper are summarized in three-folds:

(1) To the best of our knowledge, it is the first attempt to present a new method of spontaneous SER by integrating deep multimodal audio feature leaning with 1D, 2D, and 3D CNNs. Multiple high-level feature representations from three proposed CNNs are extracted as deep features, followed by an average-pooling and a score-level fusion network used for final emotion classification.
(2) We generate three different audio inputs for multi-CNNs from the original 1D raw audio waveform so as to learn deep multimodal features. Specially, inspired by 3D motions of videos, we generate appropriate 3D audio signals from the original 1D audio waveform as inputs of 3D CNNs for spatial-temporal feature learning.

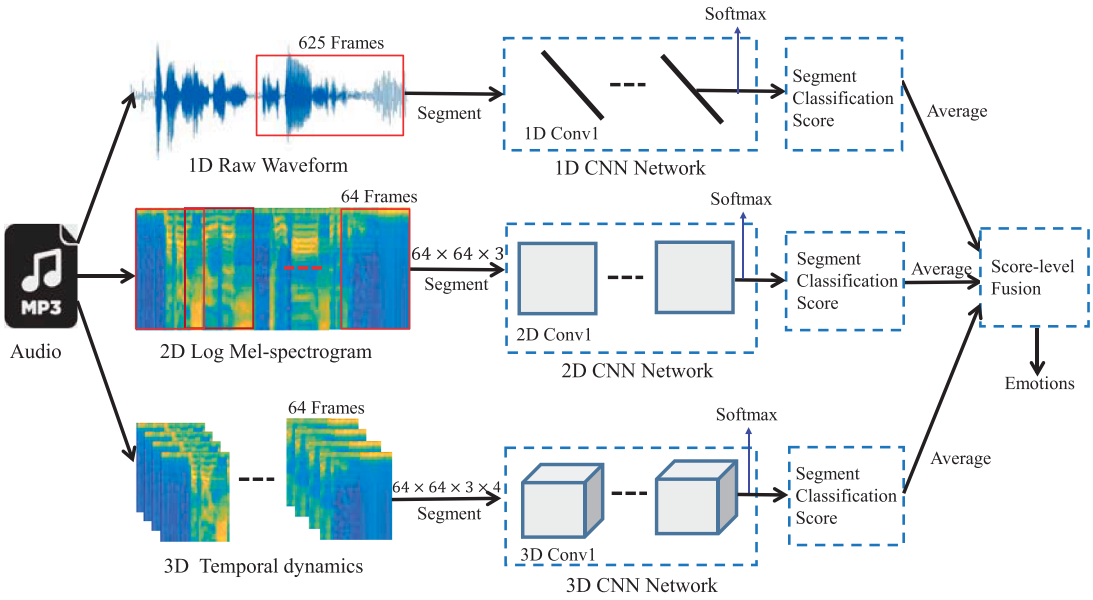

Fig. 1. Overall architecture of our proposed deep multimodal feature learning with multi-CNNs for spontaneous SER.

This is similar to video processing with 3D CNNs in computer vision.

(3) We conduct extensive experiments on two challenging spontaneous emotional speech datasets, *i.e.*, the AFEW5.0 (Dhall et al., 2015) and BAUM-1 s (Zhalehpour et al., 2017) databases. Experiment results show the validity of our proposed method on spontaneous SER tasks.

The rest of this paper is structured as follows. In Section 2 we introduce our proposed method in detail. In Section 3 we provide experiment results and analysis. Finally, we present conclusions and future work in Section 4.

2. Proposed method

Fig. 1 presents the overall architecture of our proposed deep multimodal feature learning with multi-CNNs for spontaneous speech emotion recognition. Our proposed method comprises of three steps: (1) Generating appropriate multimodal audio representations, (2) learning multimodal audio features with multi-CNNs, (3) fusing multimodal results at score-level. In the followings, we will describe the details of the above-mentioned three steps.

2.1. Generating appropriate multimodal audio representations

We generate three appropriate audio representations from the original 1D audio waveform, as described below.

(1) The first generated audio representation is the segment-level 1D raw audio waveform as an input of a 1D CNN network. The sample-level length is set to 625 frames for its best performance, which is verified in the following experiments. The raw audio waveforms are sampled at 22 kHz, and scaled into [−256, 256]. In this case, the scaled data are naturally near zero so that they are not needed to subtract the mean.

(2) Compared with 1D raw waveform, it is more popular to learn deep features from 2D time-frequency representations (*e.g.*, spectrograms) as inputs of 2D CNNs. In particular, an audio sample is divided into fixed-length segments, followed by feature learning with 2D CNNs, segment-level emotion classification, and utterance-level aggregation (Huang and Narayanan, 2017; Mao et al., 2014; Trigeorgis et al., 2016; Zhang et al., 2018b; Zhao et al., 2019b, 2018). After feature extraction, an audio sample is represented by a set of segment-level time-frequency representations that can be directly used as 2D images to conduct feature learning tasks with 2D CNNs.

In this work, we generate segment-level based 2D log Mel-spectrograms with size of $64 \times 64 \times 3$ as inputs of 2D CNNs. In detail, for an utterance we utilize 64 Mel-filter banks spanning from 20 Hz to 8000 Hz to extract the entire log Mel-spectrogram, as done in (Zhang et al., 2018c). A Hamming window size of 25 ms with an overlap of 10 ms is used. Then, a context window of 64 frames is adopted to split the extracted entire log Mel-spectrogram into fixed-length segments with size of 64×64. This is the produced static segment of 64×64. In this case, the duration of each segment hence is $10 \text{ ms} \times 63 + 25 \text{ ms} = 655$ ms. Subsequently, we calculate the first order and second order regression coefficients of the produced static segment along the time axis. Consequently, we can obtain the delta and delta-delta coefficients of the produced static segment. Finally, three channels (static, delta, and delta-delta) of Mel-spectrogram segments can be created with size of $64 \times 64 \times 3$, which is similar to the color RGB images in computer vision.

(3) It is noted that 1D raw audio waveforms and 2D time-frequency Mel-spectrograms cannot be used for temporal-spatial feature learning. To address this issue, we extract a sequence of 2D time-frequency Mel-spectrogram segments to model an audio sample as a video-like "3D temporal dynamics" for temporal-spatial feature learning. To this end, on the basis of the obtained three channels of Mel-spectrogram segments, we directly concatenate 4 successive Mel-spectrogram segments to create a video-like segment $64 \times 64 \times 3$ as inputs of 3D CNNs. Note that we use 4 successive Mel-spectrogram segments to produce video segments due to its best performance.

2.2. Learning multimodal audio features with multi-CNNs

(1) 1D raw waveform modeling with 1D CNNs

Considering the good property of sample-level CNNs, we follow this path and propose a 1D deep convolutional network including four one-dimensional convolutional layers (Conv1, Conv2, Conv3, and Conv4), followed by batch normalization (BN), rectified linear unit (ReLU), three

max-pooling layers (Pool1, Pool2, and Pool3), two full connected layers (FC5, FC6) and one softmax output layer. The stride length in convolutional layers and max-pooling layers is set to 1. The proposed 1D CNN network architecture configuration is illustrated in Table 1. Note that the output of softmax layer corresponds to the number of emotion categories on the used emotional datasets.

(2) 2D Time-frequency Mel-spectrogram modeling with 2D CNNs

Since the created three channels of Mel-spectrogram segments are similar to the RGB images, it is possible to perform a cross-media transfer learning strategy from actual image data to audio data. To this end, we fine-tune existing pre-trained deep CNN models on target data for transfer learning. As shown in (Campos et al., 2017; Ren et al., 2017), fine-tuning is widely used for transfer learning in computer vision and relieve the problem of data insufficiency. More specially, we adopt target emotional audio data to fine-tune the well-known AlexNet (Krizhevsky et al., 2012) model, which is pre-trained on the large-scale ImageNet data in 2012. We also compare other deep CNN models which have deeper network structures than AlexNet in the following experiments. Since the fixed input size of AlexNet is 227 × 227 × 3, we resize the created three channels of Mel-spectrogram segments 64 ×64 ×3 into 227 × 227 × 3as inputs of 2D CNNs with a bilinear interpolation.

As depicted in Table 2, the used AlexNet (Krizhevsky et al., 2012) for fine-tuning comprises of five convolution layers (Conv1, Conv2, Conv3, Conv4, and Conv5), three max-pooling layers (Pool1, Pool2, and Pool5), and two fully connected layers (FC6, FC7). For a fine-tuning task, we first copy the whole network parameters from the pre-trained AlexNet to initialize the used 2D CNN network. Then, we replace the softmax output layer in AlexNet with a new label vector corresponding to the number of emotion categories used in our datasets. Finally, we retrain the used 2D CNN network with the standard back propagation strategy.

(3) temporal-spatial dynamic modeling with 3D CNNs

Based on the extracted video-like segment with size of 64 × 64 × 3 × 4, we propose a 3D-CNN network to perform temporal-spatial feature learning tasks. Table 3 shows the proposed 3D-CNN network architecture. As shown in Table 3, the proposed 3D-CNN network consists of two 3D convolutional layers (Conv1, Conv2), followed by batch normalization (BN), rectified linear unit (ReLU), two 3D max-pooling layers (Pool1, Pool2), two full connected layers (FC3, FC4) and one softmax output layer. The dropout regularization is employed to avoid the overfitting.

It is noted that we do not fine-tune existing 3D-CNN models pre-trained on video action recognition datasets (Tran et al., 2015), since the number of the extracted video-like segments from an utterance is definitely small so that fine-tuning existing 3D-CNN models on target emotional datasets does not perform better than the above-mentioned designed 3D-CNN network.

Table 1
The proposed 1D CNN architecture configuration for feature learning from 1D raw waveform.

Layers	No. of Filters	Filter Size
Conv1	32	[6, 1]
Pool1	32	[5, 1]
Conv2	32	[5, 1]
Pool2	32	[5, 1]
Conv3	64	[5, 1]
Pool3	64	[5, 1]
Conv4	64	[4, 1]
FC5	2048	–
FC6	1024	–
Softmax	Emotion categories	–

Table 2
The used ALEXNET 2D-CNN architecture configuration for feature learning from RGB-like Mel-spectrograms.

Layers	No. of Filters	Filter Size
Conv1	96	[11, 11]
Pool1	96	[3, 3]
Conv2	256	[5, 5]
Pool2	256	[3, 3]
Conv3	384	[3, 3]
Conv4	384	[3, 3]
Conv5	256	[3, 3]
Pool5	256	[3, 3]
FC6	4096	–
FC7	4096	–
Softmax	Emotion categories	–

Table 3
The proposed 3D CNN architecture configuration for temporal-spatial feature learning from 3D temporal dynamics.

Layers	No. of Filters	Filter Size
Conv1	32	[8, 8, 4]
Pool1	32	[3, 3, 1]
Conv2	32	[5, 5, 1]
Pool2	32	[3, 3, 1]
FC3	256	–
FC4	256	–
Softmax	Emotion categories	–

2.3. Fusing multimodal results at score-level

Since we train 1D, 2D, and 3D CNN networks on divided audio segments, in which the segment label equals to the corresponding utterance label, it is desired to aggregate segment-level results in an utterance into utterance-level results. To this end, we employ an average-pooling strategy to implement an average operation on all divided segment-level classification scores, producing fixed-length utterance-level classification scores. Then, these fixed-length utterance-level classification scores are maximized to achieve the corresponding recognition results of each CNN network.

Considering the existing complementarity among the 1D raw waveforms, 2D Log Mel-spectrograms, and 3D temporal dynamics, we leverage a score-level fusion strategy to combine different utterance-level classification scores for final emotion classification. This can be expressed as

$$score^{fusion} = \lambda_1 score^{1D} + \lambda_2 score^{2D} + \lambda_3 score^{3D} \quad (1)$$

where λ_1, λ_2 and λ_3 denote the weight values for different utterance-level classification scores obtained by 1D, 2D, and 3D CNNs, and their sum is set to 1. In this work,λ_1, λ_2 and λ_3 are determined by an exhaustive search in a range of [0,1] with an interval of 0.1. Note that the optimal λ_1, λ_2 and λ_3 correspond to the best emotion classification performance obtained by score-level fusion.

3. Experiments

To verify the effectiveness of our proposed method on spontaneous SER tasks, two challenging spontaneous emotional speech datasets, *i.e.*, AFEW5.0 (Dhall et al., 2015) and BAUM-1 s (Zhalehpour et al., 2017), are employed for spontaneous SER experiments. We do not use other acted emotional speech datasets for experiments, because this work focuses on spontaneous SER rather than conventional acted SER.

3.1. Experiment settings

For training 1D, 2D, and 3D CNN networks, a mini-batch size of 30 is used for input data. The maximizing epoch number is 300. The learning

rate is 0.001. To speed up the CNN's training, an NVIDIA GTX TITAN X GPU with 12GB memory is employed. We implement spontaneous SER experiments with a subject-independent cross-validation strategy, mostly used in real sceneries. In detail, the original Train and Val-sets on the AFEW5.0 dataset are adopted for experiments. On the BAUM-1 s dataset containing 31 Turkish persons, a leave-one-subject-group-out (LOSGO) strategy with five subject groups is employed. In this way, on the BAUM-1 s dataset the mean recognition accuracies in five test-runs are reported.

It is noted that we split the extracted whole Mel-spectrogram from an audio sample into a certain number of Mel-spectrogram segments to conduct segment-level feature learning with CNNs. In this situation, we set the emotion category of each divided Mel-spectrogram segment to be the corresponding utterance-level emotion category. We do not employ any data augmentation method when training 1D, 2D, and 3D CNN networks.

3.2. Databases

AFEW5.0: AFEW5.0 (Dhall et al., 2015) is a spontaneous audiovisual emotional video database developed for emotion recognition in the wild challenge in 2015. This dataset has seven emotional categories such as anger, joy, sadness, disgust, surprise, fear, and neutral. To annotate these emotions, another three annotators are employed. AFEW5.0 is divided into three parts: the Train (723 samples), Val (383 samples), and Test (539 samples) sets. This work leverages the original Train and Val-sets for experiments, since the Test set is just open to the participants in the wild challenge. Specially, the emotional sample distribution on the Train set (723 samples) is anger (118), disgust (72), fear (77), joy (145), neutral (131), sadness (107), and surprise (73). In contrast, the emotional sample distribution on the Val-set (383 samples) is anger (64), disgust (40), fear (46), joy (63), neutral (63), sadness (61), and surprise (46).

BAUM-1s: BAUM-1 s (Zhalehpour et al., 2017) is a spontaneous audiovisual emotional video database developed in 2017. This dataset contains not only six basic emotional categories, including anger, joy, sadness, disgust, fear, surprise, but also other mental states such as unsure, thinking, concentrating, and bothered. Following in (Zhalehpour et al., 2017), we only focus on recognizing six basic emotional categories, thereby yielding a subset with 521 emotional video samples in total. In particular, the emotional sample distribution on this subset (521 samples) is anger (56), disgust (80), fear (37), joy (173), sadness (134), and surprise (41).

3.3. Network training

For training 1D, 2D, and 3D CNN networks, we solve the following minimizing problem so as to update the network parameters:

$$\min_{W,\partial} \sum_{i=1}^{N} H(\mathrm{softmax}(W \cdot \gamma(a_i;\partial)), y) \tag{2}$$

where Wdenotes the weight values of the softmax layer for the network parameters ∂. $\gamma(v_i;\partial)$ is the output of the final FC layer corresponding to input data a_i, and y_i denotes the class label vector of i-th segment. Here, the segment label$y_i$equals to be the utterance-level emotion category.H represents the softmax log-loss function, as described below:

$$H(\partial, y) = -\sum_{j=1}^{c} y_j \log(y_j) \tag{3}$$

where Crepresents the whole number of emotion categories.

3.4. Results and analysis

3.4.1. Effects of sample lengths as inputs of 1D CNN networks

As shown in (Kim et al., 2018; Lee et al., 2018), the sample lengths may have an important impact on the performance of sample-level 1D CNN networks. We thus initially investigate the performance of different sample lengths as inputs of 1D CNN networks for 1D raw waveform modeling. As done in (Kim et al., 2018; Lee et al., 2018), we present the performance of four different sample lengths (125, 625, 3125, 15,625 frames) as inputs of 1D CNN networks, and the corresponding number of convolution layers. Table 4 shows the recognition performance of four different sample lengths associated with the number of convolution layers. Note that each convolution layers is followed by a max-pooling layer, except that the final convolution layer is identical to a fully connected layer.

The results in Table 4 show that on the AFEW5.0 and BAUM-1 s databases the sample length with 625 frames performs best among four different sample lengths. In detail, our method gives an accuracy of 24.02% on the AFEW5.0 dataset, and 37.37% on the BAUM-1 s dataset, respectively. It can be observed from Table 4 that larger sample length helps performance improvement. However, large sample length does not always promote the performance. This may be attributed to the fact that larger sample length reduces the number of generated samples for training 1D CNNs. Therefore, sample-level 1D CNNs does not always boost the recognition performance while the segment length increases. That is why we set 625 frames for 1D CNN networks in Fig. 1.

3.4.2. Effects of fine-tuning different pre-trained deep models for 2D CNN networks

Due to the extracted RGB-like three channels of Mel-spectrogram segments as inputs of 2D CNN networks, it is feasible to fine-tune existing deep models pre-trained on ImageNet data. To evaluate the effectiveness of fine-tuning different pre-trained deep models, we present and compare the recognition performance of fine-tuning three typical deep models, including AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan and Zisserman, 2015), and ResNet-50 (He et al., 2016) on target emotional data. The reported results of these deep models are obtained by using the average-pooling strategy on all divided segment-level classification scores, followed by the maximizing operations for final emotion recognition results.

Table 5 presents the recognition results of fine-tuning three typical deep models like AlexNet, VGG-16, and ResNet-50. From Table 5, we can observe that AlexNet slightly performs better than VGG-16 and ResNet-50. On the AFEW5.0 database, AlexNet gives an accuracy of 29.24%, whereas VGG-16 and ResNet-50 separately present an accuracy of 28.16%, and 28.55%. On the BAUM-1 s database, AlexNet, VGG-16, and ResNet-50 provide an accuracy of 42.29%, 41.73%, and 41.97%, respectively. This shows that deeper networks like VGG-16 and ResNet-50 do not yield a significant improvement over AlexNet. This may because the used emotional datasets are very limited so that the produced audio segment-level samples are not enough to train deeper networks.

3.4.3. Effects of created video lengths as inputs of 3D CNN networks

For temporal-spatial feature learning, we concatenate multiple successive audio Mel-spectrogram segments to create a video-like segment

Table 4
Recognition accuracy (%) of different sample lengths (frames) as inputs of 1D CNN networks.

Sample length/ frames	Layers (Conv.)	AFEW5.0	BAUM-1s
125	3	23.34	34.18
625	4	24.02	37.37
3125	5	22.71	36.05
15,625	6	20.17	33.84

Table 5
Recognition accuracy (%) of fine-tuning different pre-trained deep models for 2D CNN networks.

Deep models	AFEW5.0	BAUM-1s
AlexNet	29.24	42.29
VGG-16	28.16	41.73
ResNet-50	28.55	41.97

called "3D temporal dynamics" as inputs of 3D CNNs. The created video segment length equals to the number of successive audio Mel-spectrogram segments. The video length may also significantly affect the performance of 3D CNN networks.

To evaluate the performance of different video lengths as inputs of 3D CNN networks, we provide the recognition results of four different video lengths (*i.e.*, 4, 6, 8, 10 segments). For these video lengths, 3D CNN networks have the same network architectures except for the first convolution layer, in which the depth of their filter sizes in third dimension (*i.e.*, the number of concatenated successive audio segments) equals to the corresponding video length. Table 6 gives the performance of four different video lengths (*i.e.*, 4, 6, 8, 10 segments) as well as the depth of their filter sizes in third dimension for the first convolution layer.

From Table 6, we can see that, the video length with 4 successive audio Mel-spectrogram segments obtains the best performance with an accuracy of 28.46% on the AFEW5.0 database, and 37.97% on the BAUM-1 s database, respectively. With the increase of video lengths, the performance of 3D CNN networks decrease. This may be because the number of constructed video segments is reduced for training 3D CNN networks when the video length increases.

3.4.4. Comparisons of multi-CNN fusion methods

For a multi-CNN fusion network, we present and compare two multi-CNN fusion methods: feature-level fusion and score-level fusion. After finishing training each CNN network independently, we conduct feature-level fusion and score-level fusion.

For feature-level fusion, we initially achieve utterance-level features for each individual CNN network. This can be realized by implementing an average-pooling operation on the learned segment-level features represented by the outputs of final FC layer in each individual CNN network. Then, we directly concatenate all produced utterance-level features (*i.e.*, 1024-D for 1D CNN, 4096-D for 2D CNN, and 256-D for 3D-CNN) from the corresponding 1D, 2D, and 3D CNN networks to constitute a large feature vector, followed by the linear support vector machines (SVM) for final emotion classification. For score-level fusion, we adopt Eq. (1) to perform multi-CNN fusion.

Table 7 shows the recognition results of different multi-CNN fusion methods as well as the single CNN network in which the best performance is obtained. When fusing 1D, 2D, and 3D CNN networks in score-level, the optimal weight values λ_1, λ_2 and λ_3 in Eq. (1) for score-level fusion are 0.3, 0.3, 0.4 for AFEW5.0, and 0.2, 0.5, 0.3 for BAUM-1 s, respectively. Similarily, we can also obtain the optimal weight values when fusing two of 1D, 2D, and 3D CNN networks in score-level. For instance, for 1D and 2D fusion, the optimal weight values are $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ for AFEW5.0, $\lambda_1 = 0.4$ and $\lambda_1 = 0.6$for BAUM-1 s, respectively. The results in Table 7 show the following observations: Table 8

Table 6
Recognition accuracy (%) of different created video lengths as inputs of 3D CNN networks.

Video lengths	Filter size (Conv1)	AFEW5.0	BAUM-1s
4	[8, 8, 4]	28.46	37.97
6	[8, 8, 6]	24.28	35.46
8	[8, 8, 8]	24.42	35.61
10	[8, 8, 10]	24.54	35.43

Table 7
Recognition accuracy (%) of different multi-CNN fusion methods.

Methods	AFEW5.0	BAUM-1s
1D CNNs	24.02	37.37
2D CNNs	29.24	42.29
3D CNNs	28.46	37.97
Feature-level (1D, 2D)	31.23	42.75
Feature-level (1D, 3D)	30.91	40.84
Feature-level (2D, 3D)	32.05	42.96
Feature-level (1D, 2D, 3D)	33.68	43.53
Score-level (1D, 2D)	32.46	42.90
Score-level (1D, 3D)	31.88	41.65
Score-level (2D, 3D)	34.43	43.77
Score-level (1D, 2D, 3D)	35.77	44.06

Table 8
Performance (%) measure when score-level fusion obtains an accuracy of 35.77% on the AFEW5.0 database.

Emotion	Recall	F1-score
Anger	56.25	50.17
Disgust	5.00	7.47
Fear	43.48	47.57
Joy	31.75	23.52
Sadness	32.79	36.10
Surprise	15.22	22.03
Neutral	50.79	37.71

1) 2D CNNs perform best, followed by 3D CNNs, and 1D CNNs. This shows that it is effective to employ the created RGB-like 2D time-frequency audio Mel-spectrogram segments to fine-tune existing deep image models pre-trained on ImageNet data, thereby relieving the pressure of emotional data insufficiency for training deep networks.

2) Score-level fusion gives superior performance to feature-level fusion. This demonstrates that score-level fusion is more suitable for multi-CNN fusion in this work. Table 9

3) Compared with the single CNN networks like 1D, 2D, and 3D CNNs, implementing multi-CNN fusion on both feature-level and score-level obtains better performance. This shows that the learned multi-modal deep features from 1D, 2D, and 3D CNN networks are complementary to each other, so that they are integrated in a multi-CNN fusion network to achieve significantly improved emotion classification performance. This may be attributed to different properties of learned convolutional feature representations from different audio inputs corresponding to 1D, 2D, and 3D CNNs. As in (Kim et al., 2018; Lee et al., 2018), sample-level 1D CNNs employ the raw 1D audio waveforms as inputs, and thus is capable of addressing log-scale amplitude compression and phase-invariance. When using 2D Mel-spectrogram as inputs, 2D CNNs have the ability of capturing energy modulation patterns across time and frequency (Zhang et al., 2018c). For temporal-spatial feature learning, several successive 2D Mel-spectrogram segments are concatenated to produce a video-like segment as inputs of 3D CNNs. Extracting such 3D video-like Mel-spectrogram segments aim to emphasize different spectral characteristics from neighboring regions in an utterance in the temporal and spatial dimension. In this case, 3D CNNs can effectively model the spatial-temporal dynamics of audio

Table 9
Performance (%) measure when score-level fusion obtains an accuracy of 44.06% on the BAUM-1 s database.

Emotion	Recall	F1-score
Anger	16.07	22.66
Joy	55.49	37.67
Sadness	70.90	39.14
Fear	5.41	9.65
Disgust	28.75	31.97
Surprise	4.88	8.93

signals. As a result, these CNN-based learned features are complementary to each other, thereby combining them for potential performance improvement on SER tasks.

To provide the recognition accuracy per emotion, Figs. 2 and 3 separately give the confusion matrices of recognition results, when our proposed method separately obtains an accuracy of 35.77%, and 44.06% with a score-level fusion on these two datasets. As depicted in Fig. 2, we can see that on the AFEW5.0 database three emotions, *i.e.*, "anger", "neutral" and "fear", are distinguished with an accuracy of 56.25%, 50.79% and 43.48%, respectively. The other four emotions, *i.e.*, "disgust", "joy", "sadness", and "surprise", are classified with an accuracy of less than 33%. Especially, "disgust" and "surprise" have poor performance than other emotions. This might be because that the audio cues of these two emotions are not distinct enough. In particular, "disgust" is wrongly identified as "joy" with 40%. From Fig. 3 we can observe that on the BAUM-1 s database two emotions, *i.e.*, "sadness" and "joy", are individually recognized with an accuracy of 70.90%, and 55.49%. The other four emotions, *i.e.*, "anger", "fear", "disgust", "surprise", are identified with an accuracy of less than 29%. Specially, "fear" and "surprise" are classified with the lowest accuracy of about 5%. The reason may be that these two emotions have confused audio cues with other emotions. For instance, "fear" is wrongly classified as "sadness" with 45.95%, whereas "surprise" is wrongly identified as "joy" with 43.90%.

3.4.5. Comparisons of states-of-the-arts

Now we compare our reported performance with previously published works on these two emotional databases. Note that these comparing works also use the same subject-independent cross-validation strategy as ours for experiments. We show the performance comparisons of state-of-the-art methods in Table 10. Note that we present the recognition accuracy rather than F1-score for performance comparisons, since most comparing works do not report their F1-score values as performance evaluation.

The results in Table 10 indicate that our proposed method is highly comparable to state-of-the-art methods on these two databases. This demonstrates the advantages of our proposed method over previously published works. For instance, compared with these works (Cai et al., 2019; Ebrahimi Kahou et al., 2015; Kayaoglu and Eroglu Erdem, 2015; Wu et al., 2015), in which the popular hand-designed INTERSPEECH 2010 set are used as speech features, our learned deep multimodal features with multi-CNNs show better performance. This exhibits the superiority of our learned deep multimodal features over hand-designed features. In addition, our proposed method also performs better than other CNN-learned methods (Ma et al., 2019; Zhang et al., 2018c), demonstrating the effectiveness of our method.

4. Conclusions and future work

Considering the existing complementarity among the learned multimodal feature representations from 1D, 2D, and 3D CNN networks, this paper proposes a new method of spontaneous SER by means of deep multimodal audio feature learning with multi-CNNs. The key step of our proposed method is to generate appropriate inputs for 1D, 2D, and 3D CNN networks from the original 1D audio waveforms, and design suitable CNN network architectures for multimodal feature learning. Experiment results on the popular AFEW5.0 and BAUM-1 s databases indicate that our proposed method performs better than state-of-the-art methods.

Note that we independently train 1D, 2D, and 3D CNN networks, and then conduct a score-level fusion for final emotion classification. Although such individual training strategy is simple yet effective, it is interesting to develop an advanced end-to-end learning system, in which 1D, 2D, and 3D CNN networks are integrated for training jointly. Besides, instead of the used multi-CNN fusion in feature-level and score-level, it is also interesting to investigate the performance of other advanced fusion methods such as factorized bilinear pooling (FBP) (Zhang et al., 2019b).

In addition, we adopt the typical softmax log-loss function to train each CNN network. In recent years, some advanced loss functions, such as angular loss (Wang et al., 2017), center loss (Wen et al., 2016), and island loss (Cai et al., 2018), have been developed for CNNs. It is thus also interesting to investigate the performance of integrating such advanced loss functions associated with 1D, 2D, and 3D CNN networks in future. Besides, this work employs two spontaneous emotional databases, *i.e.*, AFEW5.0 and BAUM-1 s, due to the data sparsity of existing spontaneous emotional databases. In future, we will employ larger spontaneous emotional databases to testify the performance of our proposed method further.

Author contributions

Shiqing Zhang: Writing-original draft preparation, Xin Tao and Yuelong Chuang:Experiment test, Xiaoming Zhao: Supervision, review

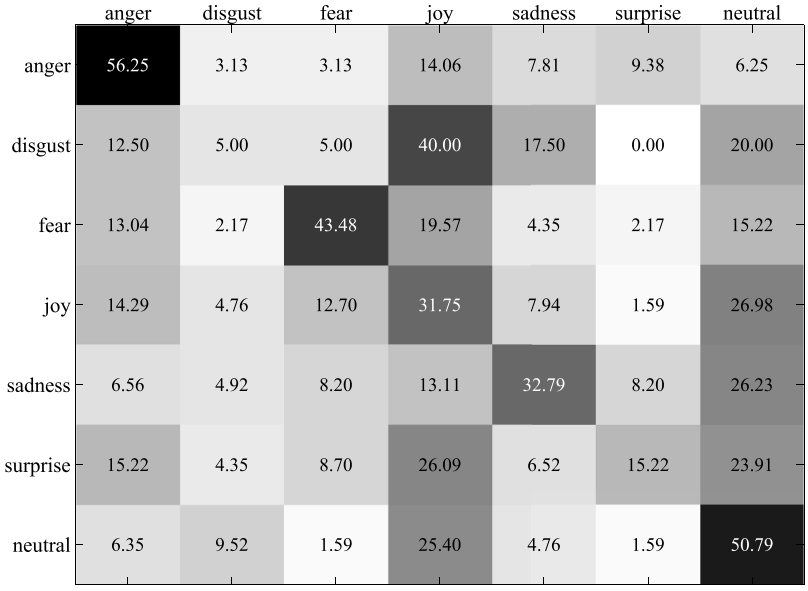

Fig. 2. Confusion matrix of recognition results when score-level fusion obtains an accuracy of 35.77% on the AFEW5.0 database.

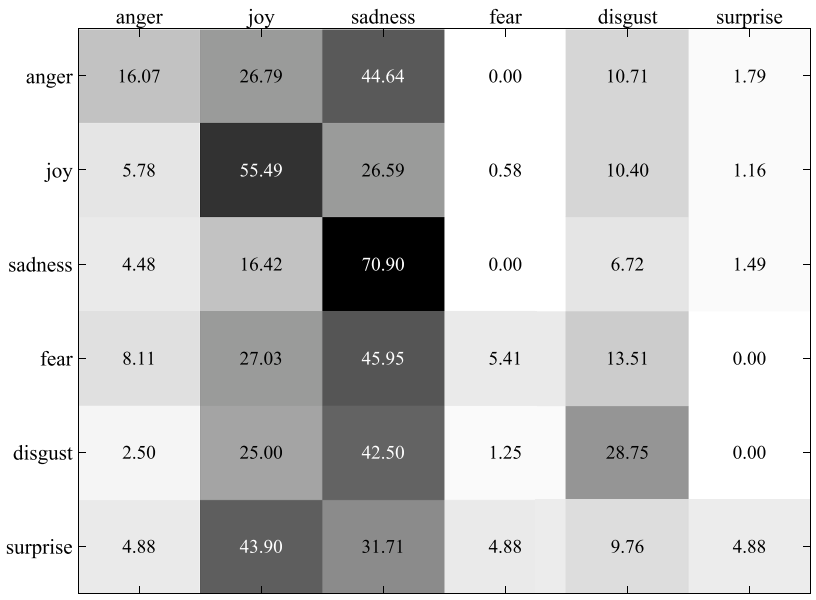

Fig.3. Confusion matrix of recognition results when score-level fusion obtains an accuracy of 44.06% on the BAUM-1 s database.

Table 10
Performance (%) comparisons of state-of-the-arts.

Datasets	Refs.	Accuracy
AFEW5.0	Kayaoglu, 2015 (Kayaoglu and Eroglu Erdem, 2015)	31.85
	Ebrahimi Kahou, 2015 (Ebrahimi Kahou et al., 2015)	33.20
	Wu, 2015 (Wu et al., 2015)	33.96
	Cai, 2019 (Cai et al., 2019)	35.51
	Ours	35.77
BAUM-1s	Zhang, 2018 (Zhang et al., 2018c)	42.46
	Zhalehpour, 2017 (Zhalehpour et al., 2017)	29.41
	Ma, 2019 (Ma et al., 2019)	42.38
	Ours	44.06

& editing.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by Zhejiang Provincial National Science Foundation of China and National Science Foundation of China (NSFC) under Grant No. LZ20F020002, LQ21F020002 and 61976149.

References

Akçay, M.B., Oğuz, K., 2020. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 116, 56–76.
Anagnostopoulos, C.-.N., Iliou, T., Giannoukos, I., 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif. Intell. Rev. 43, 155–177.
Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M.Y., Kwon, S., Baik, S.W., 2019. Deep features-based speech emotion recognition for smart affective services. Multimed. Tools Appl. 78, 5571–5589.
Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Han, S., Liu, P., Chen, M., Tong, Y., 2019. Feature-level and model-level audiovisual fusion for emotion recognition in the wild. In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval. IEEE, San Jose, USA, pp. 443–448.
Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y., 2018. Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, Xi'an, China, pp. 302–309.
Campos, V., Jou, B., Giro-i-Nieto, X., 2017. From pixels to sentiment: fine-tuning CNNs for visual sentiment prediction. Image Vis. Comput. 65, 15–22.
Demircan, S., Kahramanli, H., 2017. Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. Neural Comput. Appl. 1–8.
Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T., 2015. Video and image based emotion recognition challenges in the wild: emotiw. In: 2015, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, Seattle, pp. 423–426.
Dong, S., Gao, Z., Pirbhulal, S., Bian, G.-.B., Zhang, H., Wu, W., Li, S., 2020. IoT-based 3D convolution for video salient object detection. Neural Comput. Appl. 32, 735–746.
Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C., 2015. Recurrent neural networks for emotion recognition in video. In: ACM on International Conference on Multimodal Interaction (ICMI). ACM, pp. 467–474.
El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. 44, 572–587.
Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7, 190–202.
Fayek, H.M., Lech, M., Cavedon, L., 2017. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Netw. 92, 60–68.
Gharavian, D., Sheikhan, M., Nazerieh, A., Garoucy, S., 2012. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput. Appl. 21, 2115–2126.
Graves, A., 2012. Supervised Sequence Labelling with Recurrent Neural Networks. Springer.
Han, K., Yu, D., Tashev, I., 2014. Speech emotion recognition using deep neural network and extreme learning machine. Interspeech 223–227.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, pp. 770–778.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.
Huang, C.-.W., Narayanan, S.S., 2017. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, Hong Kong, China, pp. 583–588.
Kayaoglu, M., Eroglu Erdem, C., 2015. Affect recognition using key frame selection based on minimum sparse reconstruction. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, Seattle, WA, USA, pp. 519–524.
Kim, T., Lee, J., Nam, J., 2018. Sample-level CNN architectures for music auto-tagging using raw waveforms. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Calgary, AB, Canada, pp. 366–370.
Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25, 1106–1114.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lee, J., Park, J., Kim, K., Nam, J., 2018. Samplecnn: end-to-end deep convolutional neural networks using very small filters for music classification. Appl. Sci. 8, 150.
Liu, Z.-.T., Wu, M., Cao, W.-.H., Mao, J.-.W., Xu, J.-.P., Tan, G.-.Z., 2018. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 273, 271–280.
Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., Košir, A., 2019. Audio-visual emotion fusion (AVEF): a deep efficient weighted approach. Inf. Fusion 46, 184–192.

Mao, Q., Dong, M., Huang, Z., Zhan, Y., 2014. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimedia 16, 2203–2213.
Ren, S., He, K., Girshick, R., Zhang, X., Sun, J., 2017. Object detection networks on convolutional feature maps. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1476–1481.
Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., 2013. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. INTERSPEECH-2013, Lyon, France, pp. 148–152.
Schuller, B.W., 2018. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61, 90–99.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: ICLR-2015. San Diego, CA, USA, pp. 1–14.
Song, P., 2019. Transfer Linear Subspace Learning for Cross-corpus Speech Emotion Recognition. IEEE Trans. Affect. Comput. 10, 265–275.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497. Santiago, Chile.
Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S., 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Shanghai, China, pp. 5200–5204.
Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M., 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion Challenge. ACM, Barcelona, Spain, pp. 3–10.
Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y., 2017. Deep metric learning with angular loss. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, pp. 2593–2601.
Wang, K., Su, G., Liu, L., Wang, S., 2020. Wavelet packet analysis for speaker-independent emotion recognition. Neurocomputing. https://doi.org/10.1016/j.neucom.2020.02.085.
Wang, Z.-.Q., Tashev, I., 2017. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, New Orleans, LA, USA, pp. 5150–5154.
Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision (ECCV). Springer, Amsterdam, The Netherlands, pp. 499–515.
Wu, J., Lin, Z., Zha, H., 2015. Multiple Models Fusion for Emotion Recognition in the Wild. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, Seattle, Washington, USA, pp. 475–481.
Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E., 2017. BAUM-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans. Affect. Comput. 8, 300–313.
Zhang, B., Provost, E.M., Essl, G., 2018a. Cross-corpus acoustic emotion recognition with multi-task learning: seeking common ground while preserving differences. IEEE Trans. Affect. Comput. https://doi.org/10.1109/TAFFC.2017.2684799.
Zhang, S., Zhang, S., Huang, T., Gao, W., 2018b. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans. Multimedia 20, 1576–1590.
Zhang, S., Zhang, S., Huang, T., Gao, W., Tian, Q., 2018c. Learning affective features with a hybrid deep model for audio-visual emotion recognition. In: IEEE Transactions on Circuits and Systems for Video Technology, 28, pp. 3030–3043.
Zhang, S., Zhao, X., Tian, Q., 2019a. Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. IEEE Trans. Affect. Comput. https://doi.org/10.1109/TAFFC.2019.2947464.
Zhang, Y., Wang, Z.-.R., Du, J., 2019b. Deep fusion: an attention guided factorized bilinear pooling for audio-video emotion recognition. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, Budapest, Hungary, Hungary, pp. 1–8.
Zhao, H., Xiao, Y., Han, J., Zhang, Z., 2019a. Compact convolutional recurrent neural networks via binarization for speech emotion recognition. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Brighton, United Kingdom, pp. 6690–6694.
Zhao, X., Zhang, S., 2015. Spoken emotion recognition via locality-constrained kernel sparse representation. Neural Comput. Appl. 26, 735–744.
Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., Schuller, B., 2019b. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. IEEE Access 7, 97515–97525.
Zhao, Z., Zhao, Y., Bao, Z., Wang, H., Zhang, Z., Li, C., 2018. Deep spectrum feature representations for speech emotion recognition. In: Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data. ACM, Seoul, Republic of Korea, pp. 27–33.
Zixing, Z., Coutinho, E., Jun, D., Schuller, B., 2015. Cooperative learning and its application to emotion recognition from speech. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23, pp. 115–126.