Attention Based Convolutional Neural Network with Multi-frequency Resolution Feature for Environment Sound Classification

Minze Li\(^1\) · Wu Huang\(^2\) · Tao Zhang\(^1\)

Accepted: 15 September 2022 / Published online: 24 October 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
The environmental sound classification has great research significance in the fields of intelligent audio monitoring and other fields. A novel multi-frequency resolution (MFR) feature is proposed in this paper to solve the problem that the existing single frequency resolution time–frequency features of sound cannot effectively express the characteristics of multiple types of sound. The MFR feature is composed of three features with different frequency resolutions, which are compressed in varying degrees at the time dimension. This method not only has the effect of data augmentation but also can obtain more context information during the feature extraction. And the MFR features of Log-Mel Spectrogram, Cochleagram, and Constant Q-Transform are combined to form a multi-channel MFR feature. Also, a network named SacNet is built, which can effectively solve the problem that the time–frequency feature map of sound contains more invalid information. The basic structural unit of the SacNet consists of two parallel branches, one using depthwise separable convolution as the main feature extractor, and the other using spatial attention module to extract more effective information. Experiment results have demonstrated that the proposed method achieves the state-of-the-art accuracy of 97.5\%, 93.1\%, and 95.3\% on three benchmark datasets of ESC10, ESC50, and UrbanSound8K respectively, which are increased by 3.3\%, 0.5\%, and 2.3\% respectively compared with the previous advanced methods.

Keywords Convolutional neural network · Environment sound classification · Multi-frequency resolution · Time–frequency feature · Depthwise separable convolution · Spatial attention module

Wu Huang
huangwu@scu.edu.cn
Minze Li
243594116@qq.com
Tao Zhang
ztuestc@outlook.com

1 Chengdu Techman Sofeware Co., Ltd., Chengdu, Sichuan, China
2 Sichuan University, Chengdu, Sichuan, China
1 Introduction

Sound is not only the main carrier of information but also one of the important ways for human beings to perceive the objective world. According to the characteristics of the sound source, sounds in nature can be roughly divided into human voice and environmental sound. Accordingly, there are two main research topics of sounds: Automatic Speech Recognition (ASR) and the Environment Sound Classification (ESC). ASR converts stationary voice signals into text information and has developed relatively mature. Since ESC is required to accurately classify non-stationary acoustic signals, generally in complicated environments with various unknown sound sources including noises, it is relatively difficult and is still an open research topic ESC has been widely used in intelligent audio monitoring, scene analysis, and machine perception, etc. [1–3].

The process of ESC consists typically of sound feature extraction and classification. The feature extraction usually takes place in the time domain or the time–frequency domain. In the time domain, it deals directly with the original waveform of audio signals. The most effective method is to conduct a one-dimensional convolution on the original waveform [4]. In the time–frequency domain, the original sound signal is divided frame-by-frame and then each frame is transformed into the frequency domain by Fast Fourier Transform (FFT), thereby forming the time–frequency domain features [5]. The widely used sound feature extraction techniques are the Mel-frequency Cepstral Coefficients (MFCC) [6] and the Gammatone Frequency Cepstral Coefficients (GFCC) [7]. However, in this paper, we explore other sound feature extraction methods, such as Log-Mel Spectrogram (LM) [8], Cochleagram [9], and the Constant Q-Transform (CQT) [10]. Although the above methods have their advantages, the extracted features only have a single frequency resolution due to the fixed frame operation, which makes the expression of features have a great limitation. Because feature expression under different frequency resolutions has a great impact on sound classification [11].
Another crucial step of affecting ESC’s performance is the classification. Many machine learning algorithms have been widely used in sound classification, such as the Decision Tree classifier [12], Naive Bayes [13], Random Forest [14], Support Vector Machines (SVM) [15]. Nowadays, with the development of deep learning, the methods based on the neural networks show great advantages for ESC feature expression compared with other traditional machine learning classification algorithms. Convolutional Neural Network (CNN) is the most widely studied and many of the most advanced State of the Art (SOTA) results are also based on the research results of CNN [16–19]. Also, some ESC methods based on Recurrent Neural Network (RNN) and Capsule Networks have shown promising results [20, 21]. Although CNN-based methods have advantages for sound feature extraction and classification over other methods in ESC tasks, there is still a certain gap comparing with the outstanding performance of CNN in the field of computer vision. The time–frequency feature map in ESC is very different from the natural image, which has the characteristics of small inter-class differences and more interference information. Therefore, an appropriate neural network model is vital for ESC according to the characteristics of the time–frequency feature map, to achieve higher classification accuracy.

In this paper, a novel multi-frequency resolution (MFR) feature is proposed to address the limitation of single frequency resolution feature representation. The MFR features of LM, Cochleagram, and CQT are combined to form a multi-channel MFR feature. For MFR features, we design a network named Spatial attention convolution Network (SacNet), in which the Depth Convolution (DC) module uses depthwise convolution as the main feature extractor, and the Spatial Attention (SA) module uses the attention mechanism to extract more effective information. Using these techniques allows our model to achieve state-of-the-art performance on all the three benchmark datasets for the ESC task, namely, ESC10 [22], ESC50 [22], and UrbanSound8K [23].

The remainder of the paper is organized as follows: Sect. 2 briefly enlists the researches on deep learning for ESC tasks. Section 3 details the MFR features and the SacNet networks using multi-channel MFR features as input; Sect. 4 is the experimental section, including the introduction of the dataset and experimental setup, the comparison between the proposed method and the state-of-the-art method, and the analysis of the selection of major parameters. Section 11 draws some conclusions and presents future work.

2 Related Works

In ESC tasks, many high-performance methods based on deep learning and traditional machine learning have been proposed. In this paper, we will focus only on the method based on the neural network model and the three benchmark datasets mentioned above. In the literature, the research direction can be roughly divided into two aspects according to the dimension of audio characteristics. On the one hand, the one-dimensional (1D) time-domain features of the original waveform are extracted and classified directly, on the other hand, different two-dimensional (2D) time–frequency features are taken into account.

The most prominent characteristics of using the original waveform feature directly in ESC tasks are the low number of parameters and no need to design 2D artificial features. A large number of studies have proved the validity of the 1D time-domain waveform feature of sound. Dai et al. [4] proposed a very deep CNN, which directly uses the time-domain waveform as the input of the network. The network adopts the strategy of residual learning and the fully convolutional network structure is designed to maximize the learning ability of
the network. To simulate a bandpass filter, a convolution of large receptive fields is used at the initial layer of the network. The performance of this method in ESC tasks is sufficient to match the characteristics of 2D LM at that time. An innovative method combining vision and sound to synchronously learn sound representation was proposed by Ayter et al. [16]. The model is called SoundNet, which uses a large number of unmarked videos with audio to learn the rich natural sound representation. Unmarked videos that can be economically captured on a large scale act as bridges and the SoundNet translates discriminative visual knowledge from mature visual recognition models into sound modes. Finally, the output layer of this model is removed and used as a feature extractor to train an SVM classifier. Tokozume et al. [24] first proposed an end-to-end ESC system named EnvNet, which achieved competitive performance on three benchmark datasets. After that, EnvNetv2 [25] was proposed, which used a novel Between-Class (BC) learning method to learn to distinguish feature spaces. Two different sounds are mixed at a random rate to generate the interclass sound and input into EnvNetv2 to output the mixing ratio. BC learning method can not only improve the performance of EnvNetv2, but also other ESC models. Another novel method combining CNN and RNN was proposed by Sang et al. [20] for ESC tasks. Two different convolution modes are used in the same ESC model. CNN first performs standard feature extraction on the original sound waveform, and then RNN is used to extract time-gathering features. The results on UrbanSound8K show the effectiveness of this approach. Abdoli et al. [26] proposed an end-to-end ESC model based on CNN. This model uses a sliding window to split the signal into overlapping frames, allowing it to process audio signals of any length and showing better performance on UrbanSound8K than any other 1D feature approach. Chong et al. [27] proposed an MC-DCNNs network with features of multi-layer and multi-scale fusion. The original sound waveform was input into three branch networks with different scales, then the last three layers of the three branch networks were used for feature fusion and finally, the features were further extracted by global average pooling. This method also shows competitive performance on three benchmark datasets. This method of using 1D convolutional neural network to directly extract waveform features has also been successfully applied to COVID-19 disease diagnosis [28, 29].

CNN has made remarkable achievements in natural image classification. When the 1D waveform feature of sound is changed into the 2D time–frequency features by different methods, it not only has more feature information but also maximizes the feature extraction capability of CNN. Therefore, a lot of research works on ESC tasks are based on the time–frequency features of two dimensions. Some well-known state-of-the-art image classification networks, such as AlexNet [30] and GoogleNet [31], have been directly used in ESC tasks by Boddapati et al. [17]. MFCC and other time–frequency feature maps are sent into the network as input images, and AlexNet and GoogleNet have achieved encouraging results on the three benchmark datasets. Tang et al. [32] proposed an improved CNN named AecNet for ESC tasks. To better extract the dynamic features of sound fragments, the author adopts the multi-channel spectral features composed of LM features, first-order and second-order increments along frequency and time. AecNet extracts feature at different levels to retain low-level and high-level feature information and shows good performance on multiple datasets. In the field of image segmentation, dilated convolution is introduced into ESC tasks by Chen et al. [33]. The authors explored the effects of different expansion rates and the number of dilated convolutional layers on ESC and demonstrated in UrbanSound8K that dilated CNN have higher performance than CNN using maximum pooling and other latest methods. Zhang et al. [18] proposed a lightweight dilated CNN named LD-CNN. To reduce the number of parameters in the convolutional layer, the standard 2D convolution filter (L × W) is decomposed into two separable 1D convolution filters (L × 1 and 1 × W), and the final full connection layer
is replaced by the feature sum layer. These creative operations enable models with fewer parameters to achieve higher classification accuracy.

To integrate the advantages of a 1D waveform feature and 2D time–frequency feature of sound, some researchers used different fusion methods. Zhu et al. [34] proposed an ESC model named WaveMsNet, which first extracted the multi-scale features of the original waveform using a 1D convolution of different sizes and fused with LM features. 2D CNN is used to extract features and classify sounds. This fusion approach results in a significant improvement in ESC performance. A decision-level fusion method is used by Li et al. [35] in ESC tasks. The original waveform feature, LM, and gradient feature were input into two different networks to obtain the category score, and then the final classification result was obtained by using the Dempster-Shafer (DS) evidence theory [36] for fusion. The same fusion method was adopted by Su et al., and a model named TSCNN-DS was proposed [37]. In this method, a new 2D feature map was obtained by stitching multiple time–frequency features up and down, and two parallel networks were used to extract the features and classify them respectively. Finally, DS evidence theory was used to obtain the classification results. This approach of building new feature maps and multi-network fusion has produced amazing results on UrbanSound8K.

To further improve the performance of ESC, some studies have improved in detail, such as different data augmentation methods [38–41] and network attention mechanisms [19, 42]. The very innovative approaches mentioned above have greatly promoted the development of ESC. From these studies, we can know that the performance of the 2D time–frequency feature is higher than the 1D original waveform feature. Because 2D time–frequency features have more characteristic information, it is crucial to explore a more comprehensive characteristic expression of sound. Not only that, some data augmentation methods and clever network structure designs are also necessary to improve ESC performance.

### 3 Proposed Methods

#### 3.1 Multi-frequency Resolution (MFR) Feature

2D time–frequency feature extraction has a great impact on the performance of ESC, so we deeply explore the time–frequency feature extraction method of sound. For example, Log-Mel Spectrogram (LM) feature extraction method uses the Mel filter bank to map the linear spectrum to the Mel nonlinear spectrum to simulate the human auditory perception system. Mel frequency is a nonlinear frequency determined by the human ear’s sensory judgment of isometric pitch transformation, and its relationship with the commonly used linear frequency is

$$F(k) = 2595 \times \log_{10}(1 + f(x_k)/700),$$  \hspace{1cm} (1)

where $x_k$ represents the original sound signal, and $F(k)$ represents the Mel frequency. The Mel filter bank divides the corresponding frequency domain signal so that each frequency segment is mapped to a constant value. Now suppose the Mel filter bank is $F_m(k)$, then the LM feature extraction formula of the $i$th frame is:

$$LM_i(N) = \log \left( \sum_{m=1}^{Z} F_m(k) \left( \sum_{n=0}^{N-1} x_i(n)e^{-j2\pi kn/N} \right)^2 \right), \hspace{1cm} 0 \leq k, n \leq N - 1,$$  \hspace{1cm} (2)

where $N$ is the frame length, which also represents a period in the Fourier transform. $Z$ is the number of Mel filters, $n$ and $k$ are the abscissas of a series of discrete points. The
original sound signal \( x_i(n) \) is first subjected to FFT to obtain the amplitude spectrum, and then multiplied and accumulated by \( Z \) filters respectively. Finally, the LM feature of the \( i \)th frame can be obtained by the logarithm operation. If a sound signal is divided into \( T \) frames, then the LM features of the entire sound signal are:

\[
LM(N) = \{LM_i(N)\} \quad i = 1, 2, ..., T.
\]

Framing is an essential step to obtain time–frequency features. When the frame length is \( N \), the frequency resolution \( R \) after conversing to time–frequency features can be expressed as:

\[
R \sim \frac{1}{\Delta f}, \quad \Delta f = \frac{f_s}{N},
\]

where \( f_s \) is the sampling frequency and \( \Delta f \) is the minimum frequency interval (that is, the smaller the interval, the higher the frequency resolution). It can be seen from the above formula that when the frame length is fixed, the frequency resolution is also a fixed value. This fixed framing operation makes the extracted features only have a single frequency resolution. When the length of the sound signal is constant, the frequency resolution increases as the frame becomes longer. However, the frame number \( T \) decreases accordingly so that the time resolution decreases. It is not that the higher the frequency resolution obtained, the better results. It can be seen from Table 5 in Sect. 4.4 that the optimal frequency resolution can be obtained through experiments, but this can only show that all types of sound can reach the overall optimal at this frequency resolution. However, different sounds have the different performance under different frequency resolutions. It is obvious from Fig. 5 in Sect. 4.4 that the optimal frequency resolution of different sounds is not the same. Moreover, the single frequency resolution feature map makes the expression of features have great limitations. In summary, we propose a novel feature extraction method that can take into account multiple frequency resolutions, to obtain a more comprehensive MFR feature expression.

The proposed MFR feature is a combination of three features with different frequency resolutions. Three different frame lengths are used in the LM feature extraction to obtain three different frequency resolution feature maps denoted by \( LM(R_1) \), \( LM(R_2) \), and \( LM(R_3) \) with dimensions of \( Z \times T_1 \), \( Z \times T_2 \), and \( Z \times T_3 \). The three feature maps can be stitched together to obtain multi-frequency resolution features \( LM(MFR) \), as follows:

\[
LM(MFR) = LM(R_1) \oplus LM(R_2) \oplus LM(R_3)
\]

where "\( \oplus \)" in the formula means that the three feature maps are stitched in the time dimension, and the dimension of the \( LM(MFR) \) after stitching is \( Z \times (T_1+T_2+T_3) \). We also explore the characteristics of cochlea and CQT. The MFR features are extracted in the same way to form \( Cochl(MFR) \) and \( CQT(MFR) \) features, formulated respectively as

\[
Cochl(MFR) = Cochl(R_1) \oplus Cochl(R_2) \oplus Cochl(R_3)
\]

\[
CQT(MFR) = CQT(R_1) \oplus CQT(R_2) \oplus CQT(R_3)
\]

Figure 1 is a schematic diagram of \( LM(MFR) \) features. We set \( N = 1024, 2048, \) and \( 4096 \) to obtain the three frequency resolution feature maps of the dog barking in UrbanSound8k, which can be stitched to obtain the \( LM(MFR) \) feature. \( LM(MFR) \) is divided into three areas, and the three frequency resolution features are indicated by the dashed red, green, and orange dotted boxes. The three frequency resolution feature maps have different sizes in the time dimension, so after stitching from three sizes of regions. As can be seen from the figure, \( Area2 \) and \( Area3 \) can be regarded as different degrees of compression of \( Area1 \) in the
time dimension, similar to the fast-play effect of the original sound. This method not only has the merit of data augmentation but also extracts more context information during feature extraction. The three positions of Loc.1, Loc.2, and Loc.3 are marked as different ranges of a $k \times k$ convolution kernel extraction feature. Loc.2 and Loc.3 contain more information than Loc.1. in the time–frequency dimension. The associated information of the sound signal in the time and frequency dimensions is extremely important for the characterization of sound. The $LM(M FR)$ feature is particularly prominent at this point. It follows that benefited from the multiple frequency resolution features, the proposed MFR feature has a stronger ability to represent time–frequency dimensional correlation information.

3.2 ESC Model with Spatial Attention Convolution Network (SacNet)

The way of fusing multiple features has been proven to achieve higher performance in the ESC task, so we use three different MFR features, which are $LM(M FR)$, $Cochl(M FR)$ and $CQT(M FR)$. Unlike other fusion methods, we propose to combine three MFR features into three feature channels and as the input of the network. In particular, the proposed ESC model is shown in Fig. 2. First, the original waveform features of the sound are transformed into three MFR features. Then the three-channel features are input into the proposed SacNet network to extract features and identify the sound category. The main structural unit in the SacNet is the Spatial attention convolution block (Sac block), and see Fig. 3 for its structure. The Sac block consists of the Depth Convolution (DC) module and the proposed Spatial Attention (SA) module.

The DC module is composed of pointwise convolution and depthwise convolution. Pointwise convolution is a $1 \times 1$ convolution, which is located in the first and third layers of the DC module and is used for the expansion and compression of the number of feature channels because the feature extraction in the expanded high-dimensional feature map is more conducive to the feature expression. Depthwise convolution is different from the standard convolution, which is used in MobileNetV2 [43] as a more efficient convolution filter. Here the feature channel is expanded by 6 times, and the depthwise convolution is used to...
extract sound features in high-dimensional feature maps. Both the first and second layers use Batch Normalization (BN) and ReLU6 activation functions. In the third layer of feature compression, ReLU6 is discarded to prevent the loss of a large number of useful features.

In the SA, the feature map of sound usually contains detailed information about the foreground and complex background. The lower the signal-to-noise ratio (SNR) of the sound, the more complicated the background information, which brings huge interference to the effective information extraction of the sound. Therefore, we design an SA module during the feature extraction process in low SNR scenarios. This structure reduces the proportion of the extracted background information layer by layer, focusing more on the foreground area, which helps to generate effective features for ESC. To extract the correlation information in the time–frequency dimension more effectively without adding parameters, the SA module uses $1 \times k$ and $k \times 1$ convolutions to focus on the correlation information of the time dimension and frequency dimension, respectively. The spatial feature map after encoding is scaled to $[0,1]$ through the Sigmoid operation.

The low-level and high-level features are defined as $X \in \mathbb{R}^{H \times W \times C}$ and $X' \in \mathbb{R}^{H \times W \times C'}$ respectively, where $C$ and $C'$ are the number of feature channels. The set of spatial locations is denoted as $\mathcal{R} = \{(x, y) | x = 1, 2, ..., W; y = 1, 2, ..., H\}$, where $(x, y)$ is the spatial coordinate of features. Sac block can be described as:

$$X_1 = \text{Conv}_{1 \times k}(\text{Conv}_{k \times 1}(X, W_1^H), W_1^{H^T}),$$

(8)
Attention Based Convolutional Neural Network with Multi-frequency …

\[
X_2 = \text{Conv}_{k \times 1} (\text{Conv}_{1 \times k}(X, W_2), W_2^{II}),
\]

\[
SA = \sigma(X_1 + X_2),
\]

\[
X' = SA \cdot \text{Conv}_1 \times 1 (Dw\text{Conv}_{k \times k}(\text{Conv}_{1 \times 1}(X, W_3^{I}), W_3^{II}), W_3^{III}),
\]

where \( W_1 \) and \( W_2 \) denote the weights of the two branches of SA module, \( W_3 \) denotes the weights of DC module. \( W_1^{I}, W_2^{II}, W_3^{III} \) respectively represent the weights of adjacent convolutional layers in the branch. \( \sigma \) denotes the sigmoid operation, \( Dw\text{Conv} \) denotes the depthwise convolution, and \( k \) denotes the kernel size (we set \( k = 3 \) in experiments).

SacNet’s network structure configuration is shown in Table 1. We change the number of Sac blocks and design 5 different depths of SacNet. The three-channel MFR feature map is used as the input, and the output uses global average pooling and two fully connected layers, where a probability of 0.3 is used to randomly discard weights to prevent model overfitting. The last Softmax layer is used for category probability output.

\[\begin{array}{|c|c|c|c|c|}
\hline
\text{SacNet configuration} & \text{SacNet-6} & \text{SacNet-7} & \text{SacNet-8} & \text{SacNet-9} & \text{SacNet-10} \\
\hline
\text{Input}(H \times W \times 3) & & & & & \\
\hline
\text{Sac bolck1} & \text{Sac bolck1} & \text{Sac bolck1} & \text{Sac bolck1} & \text{Sac bolck1} & \\
C' = 32 & C' = 32 & C' = 32 & C' = 32 & C' = 32 & \\
\text{Sac bolck1-1} & & & & & \\
C' = 32 & & & & & \\
\hline
\text{Maxpool} & & & & & \\
\hline
\text{Sac bolck2} & \text{Sac bolck2} & \text{Sac bolck2} & \text{Sac bolck2} & \text{Sac bolck2} & \\
C' = 64 & C' = 64 & C' = 64 & C' = 64 & C' = 64 & \\
\text{Sac bolck2-1} & & & & & \\
C' = 64 & & & & & \\
\hline
\text{Maxpool} & & & & & \\
\hline
\text{Sac bolck3} & \text{Sac bolck3} & \text{Sac bolck3} & \text{Sac bolck3} & \text{Sac bolck3} & \\
C' = 128 & C' = 128 & C' = 128 & C' = 128 & C' = 128 & \\
\text{Sac bolck3-1} & & & & & \\
C' = 128 & & & & & \\
\hline
\text{Maxpool} & & & & & \\
\hline
\text{Sac bolck4} & \text{Sac bolck4} & \text{Sac bolck4} & \text{Sac bolck4} & \text{Sac bolck4} & \\
C' = 256 & C' = 256 & C' = 256 & C' = 256 & C' = 256 & \\
\text{Sac bolck4-1} & & & & & \\
C' = 256 & & & & & \\
\hline
\text{Global AvgPool} & & & & & \\
\hline
\text{FC-512 (Dropout = 0.3)} & & & & & \\
\hline
\text{FC-n} & & & & & \\
\hline
\text{Softmax} & & & & & \\
\hline
\end{array}\]

The depth of the configurations increases from left (6) to the right (10), the added layers are shown in bold, where \( C' \) is the number of channels and \( n \) is the sound classes.
4 Experiment Results

4.1 Datasets and Experimental Setup

To train and verify the proposed ESC model, we used ESC10, ESC50, and UrbanSound8K datasets, which are the benchmark datasets in ESC tasks.

ESC50 is composed of 50 environmental sound events of the equalization category. It contains a total of 2000 audio files, and the duration of each audio is 5 s. ESC10 is a subset of ESC50 and contains a total of 400 audio files. All the audio files are divided into 10 categories: dog barking, rain, sea waves, baby crying, clock ticking, person sneezing, helicopter, chainsaw, rooster, and fifile crackling. Compared with the first two datasets, UrbanSound8K is a relatively large dataset. It has 8732 audio files in total, and the duration of each audio is no more than 4 s, with 10 categories of unbalanced environmental sound events: air conditioner, car horn, playing children, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. Since UrbanSound8K has a problem of category imbalance and the audio duration in this data set is inconsistent, even if there are only 10 categories, its classification is more difficult than ESC50. To ensure the consistency of the network input, we pad the audio signal of fewer than 4 s by reusing the original audio signal.

In the preprocessing stage, all the audio signals are resampled with a rate of 44.1 kHz, the number of filters is set to 64, and the frame lengths are the three fixed values used for the multi-frequency resolution feature and half of which are used as the frameshift lengths. Some effective data augmentation methods are also used in our experiments. Each sample was time-stretched by 4 factors: {0.81, 0.93, 1.07, 1.23}. Each sample was pitch-shifted by 8 values (in semitones): {−3.5, −2.5, −2, −1, 1, 2, 2.5, 3.5}. Each sample was mixed with 4 acoustic scenes: {street-workers, street-traffic, street-people, park}. Each mix was generated by the relation $z = (1 - w) \cdot x + w \cdot y$, where $x$ is the audio signal of the original sample, $y$ is the signal of the background scene, and $w$ is a weighting parameter that is chosen randomly for each mix from a uniform distribution in the range of [0.1, 0.5].

We conducted experiments on a machine with NVIDIA 1080Ti graphics processing unit (GPU), CUDA9.0, cuDNN 7.1.4 and the Random Access Memory (RAM) is 256G. In the training process, random initialization weights and offsets are used, the batch size is 32, and the initial learning rate is 0.01. We used the cross-entropy loss function, and the automatic training stop strategy is adopted, that is, when overfitting occurs, the training will automatically stop and save the model after 10 cycles.

4.2 Comparison of State-of-the-Art Results

In the experiments, we use the k-fold cross-validation scheme in all the datasets ($k = 5$ for ESC10 and ESC50, $k = 10$ for UrbanSound8K), and conduct $k$ experiments to take the average value as the final recognition rate. Using multi-frequency resolution features, the accuracy of the proposed SacNet network with 5 different depth configuration on the three data sets is shown in Table 2. It can be seen from the table that SacNet-8 obtains the best performance, so we use 8-layer SacNet as the optimal network. This table also shows that for ESC tasks, it’s not that the more layers of the network, the better. Because the data is limited, too many network layers will make the network training process underfit, and there is a drop in accuracy instead.

In Table 3, the performance of the proposed method is compared with that of some state-of-the-art methods including 1D features, 2D time–frequency features, and mixed-use methods.
Table 2 Accuracy of different SacNets on three ESC datasets

| Models       | ESC10 | ESC50 | UrbanSound8K |
|--------------|-------|-------|--------------|
| SacNet-6     | 95.4  | 91.7  | 93.8         |
| SacNet-7     | 96.6  | 92.5  | 94.5         |
| SacNet-8     | 97.5  | 93.1  | 95.3         |
| SacNet-9     | 96.9  | 92.8  | 94.6         |
| SacNet-10    | 95.2  | 90.9  | 94.1         |

Table 3 Comparison of accuracy with previous State-of-the-art ESC models

| Model                                   | ESC10 | ESC50 | UrbanSound8K |
|-----------------------------------------|-------|-------|--------------|
| Human level [22]                        | 95.7  | 81.3  | –            |
| Raw audio + logmel + DS Theory [35]     | 83.1  | 92.6  | 92.2         |
| Mixup augmentation [40]                 | 91.7  | 83.9  | 83.7         |
| EnvNet [24]                             | 86.8  | 66.4  | 66.3         |
| EnvNetv2 [25]                           | 91.4  | 84.9  | 78.3         |
| 1D CNN [26]                             | –     | –     | 89           |
| WaveMsNet [34]                          | 93.75 | 79.1  | –            |
| Dilated CNN [33]                        | –     | –     | 78           |
| AecNet [32]                             | 84.9  | 68.6  | –            |
| Capsule networks [21]                   | –     | –     | 69.5         |
| AlexNet [17]                            | 86    | 65    | 92           |
| GoogleNet [31]                          | 86    | 73    | 93           |
| Multi-stream + temporal Attention [19]  | 94.2  | 84.0  | –            |
| Channel & temporal Attention [16]       | 94.2  | 86.5  | –            |
| MC-DCNN [41]                            | 87.6  | 73.1  | 75.1         |
| MFR + AlexNet                           | 92.8  | 83.3  | 93.8         |
| MFR + GoogleNet                         | 93.2  | 89.5  | 94.2         |
| Proposed SacNet-8                       | 97.5  | 93.1  | 95.3         |

It is clear to see that the proposed method achieves the highest accuracy on ESC10, ESC50, and UrbanSound8K, which are 97.5%, 93.1%, and 95.3%, respectively. For ESC10, due to the relatively few audio files and categories, human recognition accuracy on this data set reaches 95.7%, which has never been achieved by the previous methods. However, our proposed method is 1.8% higher than human recognition accuracy and 3.3% higher than the state-of-the-art methods. There are more sound categories on ESC50, and human recognition accuracy is only 81.3%. The highest performance achieved by the previous method on this dataset is 92.6%, which was achieved by adopting a complex method of multiple network fusion. Our method not only uses an end-to-end network structure but also improves the recognition accuracy by 0.5%. UrbanSound8K is a relatively difficult data set. Previously, a very complex network can only achieve a recognition accuracy of 93%. The recognition accuracy of our method is 2.3% higher than the State-of-the-art method. It can be seen from the above that the proposed method achieves the highest accuracy on all three benchmark data sets. In
addition, in order to show the advantages of the proposed SacNet feature extraction network in model design, we replaced SacNet in the structure of Fig. 2 with the classic AlexNet [28] and GoogleNet [29]. The experimental results prove the effectiveness of SacNet for feature extraction of acoustic signals.

4.3 Effectiveness of MFR Features and SA Module

To verify the effectiveness of the MFR feature, we perform a set of comparative experiments using the proposed SacNet-8 network. When the number of feature channels is 1, the input of SacNet-8 is changed to $H \times W \times 1$. The experiment explores the performance of the MFR features and single frequency resolution features obtained under different frequency resolutions when the window lengths are 1024, 2048, and 4096, respectively. The experimental results are shown in Table 4. As can be seen from the table, the recognition accuracy of the three MFR features based on the LM feature, Cochlf feature, and CQT feature is significantly improved over the corresponding single frequency resolution feature. The MFR features are not only valid for a certain data set. In our experiments, the MFR features show higher recognition accuracy on the three benchmark data sets. And the three MFR features are combined into three features channels as the optimal features and input the SacNet-8 network, which has higher recognition accuracy than single-channel MFR features. It can be seen that the multi-channel MFR feature has better characterization ability for sound.

SA module is an attention module designed to extract more effective feature information. A proper attention module can avoid acquiring too much invalid information, and capturing more effective feature details. To verify the effectiveness of the SA module, we use the multi-channel MFR feature as input and then use the SacNet removed the SA module as the network model to train and test on the three benchmark data sets. Figure 4 shows the recognition accuracy of the five SacNet with and without the SA module on the three data sets. It can be seen from the figure that the accuracy of the five SacNets without the SA

| Feature            | Window length | ESC10 | ESC50 | UrbanSound8K |
|--------------------|---------------|-------|-------|--------------|
| LM                 | 1024          | 91.3  | 89.4  | 90.3         |
|                    | 2048          | 93.8  | 90.5  | 91.6         |
|                    | 4096          | 91.3  | 89.1  | 90.5         |
| LM(MFR)            | –             | 95.3  | 91.2  | 93.5         |
| Cochlf             | 1024          | 94.7  | 89.6  | 92.1         |
|                    | 2048          | 95.6  | 90.8  | 93.6         |
|                    | 4096          | 94.0  | 88.9  | 91.5         |
| Cochlf(MFR)        | –             | 96.2  | 91.6  | 94.1         |
| CQT                | 1024          | 92.4  | 89.3  | 89.5         |
|                    | 2048          | 93.7  | 89.7  | 90.8         |
|                    | 4096          | 92.4  | 88.5  | 89.5         |
| CQT(MFR)           | –             | 94.6  | 90.5  | 92.0         |
| LM(MFR) + Cochlf(MFR) + CQT(MFR) | – | 97.5  | 93.1  | 95.3 |
module has decreased significantly. The differences presented on the three benchmark data sets are sufficient to reflect the role of the SA module in the SacNet.

4.4 Parameter Selection for Frequency Resolution

The MFR feature combines multiple frequency resolution features to more fully characterize the acoustic signal, making up for the lack of a single frequency resolution feature that can only effectively characterize certain types of acoustic signals. It involves the selection of multiple frequency resolution sizes, what we use is to change the window length to obtain different frequency resolution features. We use SacNet-8 as the training network and change its input to $H \times W \times 1$. ESC10 is used as the experimental dataset, and the LM feature is used as the input of the network to perform a set of experiments. R1, R2, R3, R4, R5, and R6 are the frequency resolutions obtained by taking different window lengths. The experiment results are shown in Table 5. The highest recognition accuracy of R4 on ESC10 is 93.8%, followed by R3 and R5. This experiment shows that at a single frequency resolution, the features of the three frequency resolutions R3, R4, and R5 are more effective.

To further clarify the ability of different frequency resolution features to characterize different types of acoustic signals, we show the recognition accuracy of each category under different frequency resolutions in Fig. 5. As one can see, the recognition accuracy of different sound signals under different frequency resolutions has a large difference. The sound signals

**Table 5** Average recognition accuracy under different frequency resolutions on ESC10

| Frequency resolution | Window length | Window shift | Accuracy (%) |
|----------------------|---------------|--------------|--------------|
| R1                   | 256 (5.8 ms)  | 128          | 88.7         |
| R2                   | 512 (11.6 ms)| 256          | 90.0         |
| R3                   | 1024 (23.2 ms)| 512          | 91.3         |
| R4                   | 2048 (46.4 ms)| 1024         | 93.8         |
| R5                   | 4096 (92.9 ms)| 2048         | 91.3         |
| R6                   | 5120 (116.1 ms)| 2560    | 85.5         |
can be more fully characterized at certain frequency resolutions, which will improve the recognition accuracy. Table 5 shows that the average recognition accuracy under the three frequency resolutions of R3, R4 and R5 is the highest. It can also be seen in Fig. 5 that most acoustic signals have higher recognition accuracy under these three frequency resolutions. Therefore, the three frequency resolutions of R3, R4, and R5 are adopted as the optimal parameters of MFR features in this paper.

5 Conclusion

Since the existing time–frequency features of sound only have a single frequency resolution, it is impossible to effectively express the characteristics of multiple types of sound. To surmount this problem, a novel multi-frequency resolution feature is developed, which can extract more context information during feature extraction, thereby getting higher accuracy for ESC. Furthermore, to remove the invalid information interference in the sound time–frequency feature map, the designed SacNet network is proposed to accurately extract the effective information. Experiment results over the three benchmark datasets of ESC10, ESC50, and UrbanSound8K, have shown that the proposed method achieves higher ESC accuracy, which run up to 97.5%, 93.1%, and 95.3%, respectively. More importantly, the proposed multi-frequency resolution feature can be applied to a data set with a larger number of sound categories to better explore its feature representation capabilities. In the task of environmental sound classification, the future challenge is still the lack of large-scale data sets. This leads to some existing methods that only have a good classification effect on the sounds in part of the data sets.

References

1. Gerstoft P, Hu Y, Patil C et al (2021) Audio scene monitoring using redundant un-localized microphone arrays. arXiv preprint arXiv:2103.01830
2. Paseddula C, Gangashetty SV (2021) Late fusion framework for acoustic scene classification using LPCC, SCMC, and log-mel band energies with deep neural networks. Appl Acoust 172:107568
3. Castiajo P, Pinheiro AP (2021) Acoustic salience in emotional voice perception and its relationship with hallucination proneness. Cogn Affect Behav Neurosci : 1–14.
4. Dai W, Dai C, Qu S et al (2017) Very deep convolutional neural networks for raw waveforms. IEEE Trans Pattern Anal Mach Intell 38(5):1037–1049
Attention Based Convolutional Neural Network with Multi-frequency … 4305

5. Shepard RN (1964) Circularity in judgments of relative pitch. J Acoust Soc Am 36(12):2346–2361
6. Sangeetha J, Hariprasad R, Subhiiksha S (2021) Analysis of machine learning algorithms for audio event classification using Mel-frequency cepstral coefficients. In: Applied Speech Processing. Academic Press, pp 175–189
7. Kumaran U, Rammohan S R, Nagarajan S M et al. (2021) Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. Int J Speech Technol, 1–12.
8. Piczak KJ (2015) Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP). IEEE, pp 1–6
9. Sharan RV, Moir TJ (2019) Acoustic event recognition using cochleagram image and convolutional neural networks. Appl Acoust 148(1):62–66
10. Singh P, Saha G, Sahidullah M (2021) Non-linear frequency warping using constant-Q transformation for speech emotion recognition. In: 2021 international conference on computer communication and informatics (ICCCI). IEEE, pp 1–6
11. Shu HY, Song Y, Zhou H (2018) Time-frequency performance study on urban sound classification with convolutional neural network. In: 2018 IEEE region 10 conference. Jeju, Korea (South), IEEE, pp 1713–1717
12. Tan LN, Yonggwan W (2016) Frequency features selection using decision tree for classification of sleep breathing sound. In: Information science and applications (ICISA), vol 376(9). Springer Singapore, pp 1375–1380
13. Grobelnik M (1999) Feature selection for unbalanced class distribution and Naive Bayes. In: Proceedings of the sixteenth international conference on machine learning (ICML), Bled, Slovenia, pp 258–267
14. Christoph HA, Julian B, Steffen L et al (2017) Nonnegative matrix factorization and random forest for classification of heart sound recordings in the spectral domain. In: Computing in cardiology conference. IEEE
15. Kobat MA, Dogan S (2021) Novel three kernelled binary pattern feature extractor based automated PCG sound classification method. Appl Acoust 179:108040
16. Aytar Y, Vondrick C, Torralba A (2019) SoundNet: learning sound representations from unlabeled video. pp 1–9. arXiv:1610.09001v1 [cs.CV]
17. Boddapati V, Petef A, Rasmusson J et al (2017) Classifying environmental sounds using image recognition networks. Procedia Comput Sci 112(8):2048–2056
18. Zhang XH, Zou YX, Wang WW (2018) LD-CNN: a lightweight dilated convolutional neural network for environmental sound classification. In: 2018 24th international conference on pattern recognition (ICPR). IEEE, pp 373–378
19. Zhang ZC, Xu SG, Zhang SQ et al (2019) Learning attentive representations for environmental sound classification. IEEE Access 7(2):130327–130339
20. Salamon J, Jacoby C, Bello JP (2014) A dataset and taxonomy for urban sound research. In: 22nd ACM international conference on multimedia. ACM, pp 1015–1018.
21. Tokozume Y, Harada T (2017) Learning environmental sounds with end-to-end convolutional neural network. In: 2017 ieee international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2721–2725
22. Tokozume Y, Ushiku Y, Harada T (2018) Learning from between-class examples for deep sound recognition. pp 1–9. arXiv:1711.10282v2 [cs.LG]
23. Abdoli S, Cardinal P, Koerich AL (2019) End-to-end environmental sound classification using a 1D convolutional neural network. Expert Syst Appl 136:252–263
24. Chong D, Zou Y, Wang W (2019) Multi-channel convolutional neural networks with multi-level feature fusion for environmental sound classification. MultiMed Model 11298(5):157–168
25. Lella KK, Pja A (2021) Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice. AIMS Public Health 8:240. https://doi.org/10.3934/publichealth.2021019
26. Lella KK, Pja A (2021) Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice. AIMS Public Health 8(2):240–264. https://doi.org/10.3934/publichealth.2021019 (PMID:34017889;PMCID:PMC8116184)
30. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25(2):1097–1105
31. Szegedy C, Liu W, Jia Y et al. (2015) Going deeper with convolutions. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 1–9
32. Tang G, Liang R, Xie Y et al. (2019) Improved convolutional neural networks for acoustic event classification. Multimed Tools Appl 78(12):15801–15816
33. Chen Y, Guo Q, Liang X et al. (2019) Environmental sound classification with dilated convolutions. Appl Acoust 148(5):123–132
34. Zhu B, Wang C, Liu F, et al. (2018) Learning environmental sounds with multi-scale convolutional neural network. In: 2018 international joint conference on neural networks (IJCNN). IEEE, pp 1–8.
35. Li S, Yao Y, Hu J et al. (2018) An ensemble stacked convolutional neural network model for environmental event sound recognition. Appl Sci 8(7):1152–1171
36. Dempster AP (2008) The Dempster–Shafer calculus for statisticians. Int J Approx Reason 48(2):365–377
37. Su Y, Zhang K, Wang JY et al. (2019) Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors 19(7):1733–1746
38. Salamon J, Bello JP (2017) Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 24(3):279–283
39. Davis N, Kumaraswamy S (2018) Environmental sound classification using deep convolutional neural networks and data augmentation In: 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS). IEEE, pp 41–45
40. Zhang Z, Xu S, Cao S et al. (2018) Deep convolutional neural network with mixup for environmental sound classification. In: Chinese conference on pattern recognition and computer vision (PRCV). Springer, pp 356–367
41. Tadanobu I, Phongtharin V, Shiqiang W et al. (2019) Shuffling and mixing data augmentation for environmental sound classification. In: Proceedings of the detection and classification of acoustic scenes and events 2019 workshop (DCASE). New York University, pp 109–103.
42. Li X, Chebiyyam V, Kirchhoff K (2019) Multi-stream network with temporal attention for environmental sound classification. pp 1–5. arXiv:1901.08608v1 [cs.SD]
43. Sandler M, Howard A, Zhu M, et al. (2018) MobileNetV2: inverted residuals and linear bottlenecks, pp 1–14. arXiv:1801.04381v4 [cs.CV]

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.