# Deep Convolutional Neural Network Combined with Concatenated Spectrogram for Environmental Sound Classification

Zhejian Chi, Ying Li, Cheng Chen
*College of Mathematics and Computer Science, Fuzhou University*
Fuzhou, China
zj_chi@qq.com, fj_liying@fzu.edu.cn, 704986041@qq.com

***Abstract*—Environmental sound classification (ESC) is an important but challenging issue. In this paper, we propose a new deep convolutional neural network, which uses concatenated spectrogram as input features, for ESC task. This concatenated spectrogram feature we adopt can increase the richness of features compared with single spectrogram. It is generated by concatenating two regular spectrograms, the Log-Mel spectrogram and the Log-Gammatone spectrogram. The network we propose uses convolutional blocks to extract and derive high-level feature images from concatenated spectrogram, and each block is composed of three convolutional layers and a pooling layer. In order to keep depth of the network and reduce numbers of parameters, we use filter with a small receptive field in each convolutional layer. Besides, we use the average pooling to keep more information. Our method was tested on ESC-50 and UrbanSound8K and achieved classification accuracy of 83.8% and 80.3%, respectively. The experimental results show that the proposed method is suitable for ESC task.**

***Keywords—Concatenated spectrogram, Deep convolutional neural network, Environmental sound classification.***

## I. INTRODUCTION

As an important branch of sound event recognition, environmental sound classification (ESC) can be applied to environment survey [1], smart home [2], scene classification [3] and robot hearing [4]. However, the composition of sound events in the environment is complicated, because the environment in which sound events occur is diverse, different events may overlap each other, and even background noise exists [5]. Therefore, an effective method should be designed for ESC task.

The traditional methods for ESC task rely on handcrafted features, which are modeled by some typical classifiers, such as KNN or SVM [5]. However, the performance of these methods cannot meet our expectations, one key factor is that traditional classifiers cannot extract features further [6]. Recently, convolutional neural network (CNN) has achieved great progress in many pattern recognition tasks, such as classification of traffic signs, pedestrian detection, and face recognition [7]. CNNs primarily used in field of visual recognition and it has been also successfully applied to speech [8], [9], music analysis [10], and daily sound detection and classification [7], [11], [12] [13]. Especially for the field of sound detection and classification, more and more CNN-based methods have been designed and achieved the state-of-the-art performance. For example, there are many CNN-based methods performing well on sound classification and detection task in DCASE community competition, and you can easily find these resources on the web.

By analyzing these well performing methods, we find that designing a CNN-based method for ESC task can be considered in two aspects, which are network architecture designing and input features selection. Although data augmentation always has a slight improvement in the final result, we are not going to use this trick in our method because it actually changes the original distribution space of data. Therefore in this paper, we propose a deep CNN-based method for ESC task, which uses concatenated spectrogram as input features of the CNN. The rest of this paper is organized as follows. The input features, concatenated spectrogram, of CNN are introduced in Sect. II. Section III provides detailed introduction of the proposed CNN architecture. Section IV presents the experiment settings, and gives experimental results. Finally, Sect.V gives the conclusion.

## II. CONCATENATED SPECTROGRAM

### A. Two Type Spectrograms

Many works have been dedicated to finding appropriate representations which can extract meaningful information from sound signals [14]. According to these works and the systems submitted in the DCASE community competition, we find Log-Mel spectrogram (LMS) has been widely used as input features of neural networks, such as CNN, and reach better performance compared with traditional methods. For example, Piczak's method uses the LMS and its delta information as the input features of CNN model, which created a huge improvement over baseline methods [7]. In addition to using the LMS as an input features, we add another spectrogram, the Log-Gammatone spectrogram (LGS). Similarity with Piczak's method which combines the LMS and its delta information as input features; we concatenate LMS with LGS as input features. We give the two type spectrograms of an alarm clock sound in Fig. 1.

Selecting LMS and LGS to form input features is based on the following reasons. First, both features have achieved good performance in the field of sound recognition. Secondly, the filter bank that generates them is different, so the spectrogram concatenated by them improves the richness of the input features.

### B. Concatenated Spectrogram Generation

In this section, we will simply introduce the progress of concatenated spectrogram generation from audio signal.

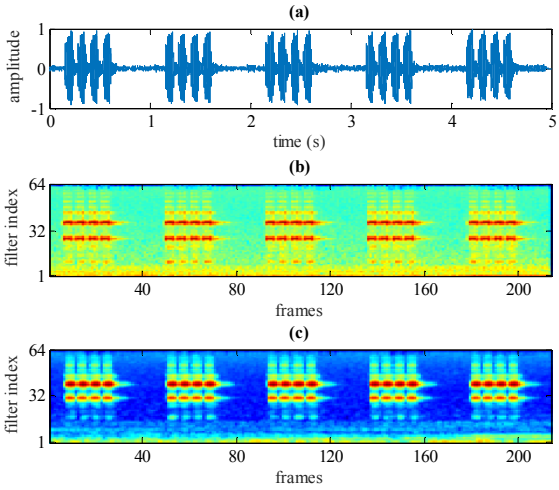


Fig. 1. Different representations of an alarm clock sound. (a) Waveform of alarm clock. (b) Log-Mel spectrogram. (c) Log-Gammatone spectrogram.

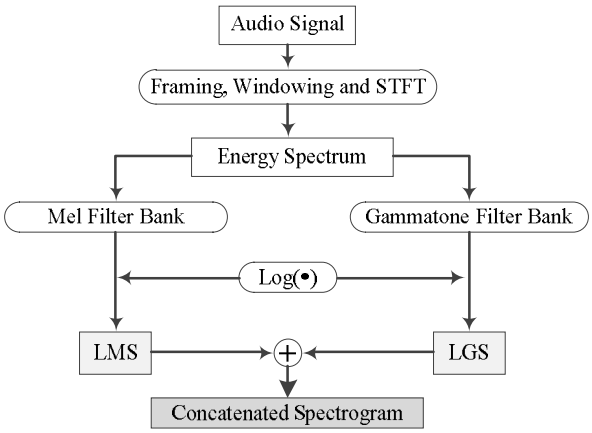


Fig. 2. The process of concatenated spectrogram generation.

Given an audio signal, we first use short-time Fourier Transform (STFT) to extract the energy spectrum. Then, we apply the Mel filter bank [15] and the Gammatone filter bank [16] to the energy spectrum respectively and the resulting spectrogram is converted into logarithmic scale. Afterwards, we concatenate the LMS and LGS to generate a joint feature representation, concatenated spectrogram, as the input features of the CNN. We summarize the feature generation process in Fig. 2.

## III. Architecture Of The Deep CNN

In this paper, we design a new deep CNN inspired by VGG net [17] for ESC task. Our CNN has the same number of weight layers and similar structure as the VGG-13 [17], but is more simplified. Our CNN structure consists of four convolutional blocks, one global max pooling layer and a fully connected (FC) layer. The detail configuration is presented in Table I.

The operations in these four convolutional blocks are the same except that the number of channels is multiplied. There are three convolutional operations in every convolutional block, among them, two convolutional operations with filter size of 3×3 are used for learning joint time-frequency patterns. In order to speed up the training and convergence of the network, batch normalization (BN) [18] is used for the output of the convolution operation. Then, we use the Rectified Linear Units (ReLU) to activate the output of BN to achieve nonlinearity of the model. In particular, we insert a convolutional layer with filter size of 1×1 without any other operation between two 3×3 convolutional layers. In general, the incorporation of 1×1 convolutional layers is a way to increase the nonlinearity of the decision function without affecting the receptive fields of the convolutional layers. But in our case, it is only a linear projection onto the space of the same dimension, which is used for reducing complexity of network and making network's depth deeper. After convolutional operation, there is an average pooling layer with filter size of 2×2 for down sampling, because the advanced semantic information after average pooling layer in the deep layers of the network can generally help the classifier to classify.

After all convolutional blocks, we use a global max pooling layer to transform the high-level feature images into vectors. The global max pooling is constant to time or frequency changing because it extracts maximum the time and frequency information in the spectrogram of the sound event. Finally, we use one FC layer on the vector followed by a softmax function which can transform output to the probability of the audio class.

## IV. Experiments and Results

### A. Dataset

We use two publicly available datasets which are ESC-50 [19] and UrbanSound8K [20] to evaluate the proposed method in this paper.

TABLE I. Configuration of the Proposed Deep CNN.

| Layer | Operation | Output Shape[a] |
|---|---|---|
| Input | - | (1,128,64) |
| Block1 | Conv3-64[b], BN[c], ReLU[d]<br>Conv1-64<br>Conv3-64, BN, ReLU | (64,128,64) |
| *Average Pooling* | (2×2)[e] | (64,64,32) |
| Block2 | Conv3-128, BN, ReLU<br>Conv1-128<br>Conv3-128, BN, ReLU | (128,64,32) |
| *Average Pooling* | (2×2) | (128,32,16) |
| Block3 | Conv3-256, BN, ReLU<br>Conv1-256<br>Conv3-256, BN, ReLU | (256,32,16) |
| *Average Pooling* | (2×2) | (256,16,8) |
| Block4 | Conv3-512, BN, ReLU<br>Conv1-512<br>Conv3-512, BN, ReLU | (512,16,8) |
| *Average Pooling* | (2×2) | (512,8,4) |
| *Global Max Pooling* | (8×4) | (512,) |
| FC | - | Number of classes |

[a] Output Shape: (channel, time, frequency).

[b] "Conv3-64" represents using 64 channels with filter size of (3×3) for convolution.

[c] BN represents Batch Normalization.

[d] ReLU is activation function.

[e] Pooling filter size.

The ESC-50 dataset contains 2000 environmental sound clips which can be roughly divided into 5 categories, including *animals*, *natural soundscapes and water sounds*, *human non-speech sounds*, *interior/domestic sounds*, *and exterior/urban noises*. Each category contains 10 classes, so there are a total of 50 classes. All audio clips are 5s with sampling of 44.1kHz , and they are pre-sorted into 5 folds.

The UrbanSound8K dataset contains 8732 labeled urban sound clips (up to 4s) from 10 classes: *air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, and street_music*. The sampling rate of all files is also 44.1 kHz, and the files are pre-sorted into 10 folds.

### B. Experiment Setup

In this paper, we calculate LMS through the LibROSA which is a python package for music and audio analysis. As for LGS, we calculate it using the method implemented by Ellis [21].

We normalized all audio samples from both datasets to a range of [-1,1]. As for two-channel audio in the dataset, we converted it to mono by averaging the two channels for building spectrogram. In the spectrogram generation phase, the Hamming window size is set to 2048 samples (sampling rate 44.1 kHz) with 50% overlap. Then, we apply Mel filter bank and Gammatone filter bank which have a cut off frequency of 50Hz on the spectrograms. The number of filters is set to 64 for both two type filter banks, which makes the generated spectrogram can be divisible in pooling layers. Finally, the logarithmic operation is performed on the spectrogram to obtain Log spectrograms.

In training phase, we use the Adam optimizer [22], the batch size is set to 64. The initial learning rate is set to 0.001, then it becomes 95% of the current learning rate after every 200 iteration. The models are trained totally 5000 iterations for the two datasets using Torch library on a Nvidia-RTX-2080 GPU with an 8 GB memory. Then, the K-fold cross-validation is adopted to evaluate the performance of proposed method. The datasets have already set the folds, so in the experiment we use these settings directly. That is, K is set to 5 and 10 for the ESC-50 and UrbanSound8K dataset, respectively.

### C. Results

*1) Performance of proposed CNN combined with different input features:* In this paper, we tested the performance of the three input features which are the concatenated spectrogram, the LMS and the LGS. The results of which are given in Table II. It can be seen that the concatenated spectrogram performs best on both datasets, which indicate that the concatenated spectrogram as input features does improve the classification effect of the proposed CNN on the environmental sound.

TABLE II. CLASSIFICATION ACCURACY(%) OF PROPOSED CNN COMBINED WITH DIFFERENT INPUT FEATURES

| Input Features | ESC-50 | UrbanSound8K |
|---|---|---|
| LMS | 81.0 | 78.0 |
| LGS | 80.1 | 77.0 |
| Concatenated Spectrogram | **83.8** | **80.3** |

TABLE III. CLASSIFICATION ACCURACY(%) BETWEEN PROPOSED CNN AND VGG-13.

| Networks | ESC-50 | UrbanSound8K | Parameters (in millions) |
|---|---|---|---|
| VGG-13 [17] | 71.5 | 71.2 | 59.95 |
| Proposed CNN | **83.8** | **80.3** | **5.44** |

TABLE IV. CLASSIFICATION ACCURACY(%) BETWEEN DIFFERENT MEHTODS

| Methods | ESC-50 | UrbanSound8K |
|---|---|---|
| Piczak [7] | 64.9 | 72.7 |
| D-CNN [15] | 68.1 | **81.9** |
| Zhang [6] | 76.8 | 74.7 |
| Proposed | **83.8** | 80.3 |
| *Human Performance* | 81.3 | - |

*2) Comparision with VGG-13:* We compare our proposed CNN with VGG-13 which has the same number of network layers. The only difference between the VGG-13 used here and the original one is that the input and output size are changed. We use the concatenated spectrogram as input features for both network. As can be seen from Table III, the proposed CNN not only performs better than the VGG-13 on both datasets, but the number of parameters in our CNN is much less than that of VGG-13.

*3) Comparision with different methods:* The previous results are all from a unilateral comparison of methods, such as features or network structure, and now we compare our method with some recent related works, which will allow a more comprehensive assessment of the performance of our approach. The results of them are listed on Table IV. It is clear that our method performs best (83.8%) on ESC-50 dataset between these methods, which obtains an absolute improvement of 18.9% comparing with Piczak's method. Although our method can not reach the best performance on UrbanSound8K, it also obtains an absolute improvement of 7.6% comparing with Piczak's method. These results show that increasing the depth of the CNN, as well as the use of filter with a small receptive field, can improve the recognition ability of the network. Futhermore, the average classification accuracy of our method also outperforms human performance on ESC-50 dataset.

## V. CONCLUSION

In this paper, we propose a deep CNN architecture inspired by VGG for ESC task. It uses the concatenated spectrogram as input features and has achieved better classification results on both the ESC-50 and UrbanSound8K dataset compared with LMS and LGS feature. Meanwhile, we test our CNN and VGG-13 on both datasets and results show that architecture of our CNN not only has better performance but has less parameters. Furthermore, we find that the proposed method can classify the environmental sound better compared with some recent CNN-based methods and has surpassed human performance on the ESC-50 dataset. In the next stage, we will continue to design CNN-based network and extract useful feature for improving the performance on ESC task.

## ACKNOWLEDGMENT

The research in this paper is supported by the Natural Science Fund Project of Fujian Province (No. 2018J01793).

## REFERENCES

[1] R. Radhakrishnan, A. Divakaran and A. Smaragdis, "Audio analysis for surveillance applications," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2005, pp. 158-161.

[2] M. Vacher, J. F. Serignat, S. Chaillol. "Sound Classification in a Smart Room Environment: an Approach using GMM and HMM Methods." The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD 2007), Publishing House of the Romanian Academy (Bucharest), Iasi, Romania. May 2007, pp.135-146.

[3] D. Barchiesi, D. Giannoulis, D. Stowell and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," in IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 16-34, May 2015.

[4] R. F. Lyon, "Machine Hearing: An Emerging Field [Exploratory DSP]," in IEEE Signal Processing Magazine, vol. 27, no. 5, pp. 131-139, Sept. 2010.

[5] C. Wang, J. Wang, A. Santoso, C. Chiang and C. Wu, "Sound Event Recognition Using Auditory-Receptive-Field Binary Pattern and Hierarchical-Diving Deep Belief Network," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 8, pp. 1336-1351, Aug. 2018.

[6] Z. Zhang, S. Xu, S. Cao, S. Zhang. "Deep convolutional neural network with mixup for environmental sound classification." Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, Cham, 2018.

[7] K. J. Piczak, "Environmental sound classification with convolutional neural networks," 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, 2015, pp. 1-6.

[8] T. N. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 8614-8618.

[9] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, Oct. 2014.

[10] A. Van den Oord, S. Dieleman, B. Schrauwen, "Deep content-based music recommendation," Advances in Neural Information Processing Systems, 2013, pp. 2643-2651.

[11] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, March 2017.

[12] I. McLoughlin, H. Zhang, Z. Xie, Y. Song and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 540-552, March 2015.

[13] X. Zhang, Y. Zou, and S. Wei, "Dilated convolution neural network with LeakyReLU for environmental sound classification." 2017 22nd International Conference on Digital Signal Processing (DSP). IEEE, 2017.

[14] Virtanen, Tuomas, M. D. Plumbley, and D. Ellis, eds. Computational analysis of sound scenes and events. Heidelberg: Springer, 2018, pp: 72-72.

[15] Kopparapu, S. Kumar, and M. Laxminarayana. "Choice of Mel filter bank in computing MFCC of a resampled speech." 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010). IEEE, 2010.

[16] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification," IEEE Transactions on Multimedia, vol. 14, no. 6, pp. 1684-1689, Dec. 2012.

[17] Simonyan, Karen, and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[18] Ioffe, Sergey, and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

[19] K. J. Piczak, "ESC: Dataset for environmental sound classification." Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015.

[20] Salamon, Justin, C. Jacoby, and J. P. Bello. "A dataset and taxonomy for urban sound research." Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.

[21] Ellis, PW. Daniel, "Gammatone-like spectrograms." web resource: http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram (2009).

[22] Kingma, P. Diederik, and Ba. Jimmy, "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).