An Environmental Sound Classification Algorithm Based on Multiscale Channel Feature Fusion

Wen Zhao[1], Helong Wang[2], Yizi Chen[3], Xuesong Pan[*,2], Kaiyue Zhang[3] and ZhenXiang Bai[1]
[1]*Ocean University of China, Qingdao, China*
[2]*Qingdao Haier Air Conditioner Co., Ltd, State Key Laboratory of Digital Household Appliances, Qingdao, China*
[3]*Qingdao Haier Air Conditioner Co., Ltd, Qingdao, China*

Keywords: Environmental Sound Classification, Feature Fusion, Deep Learning.

Abstract: In recent years, the automatic classification of urban environmental sounds has emerged as a pivotal task in the urban informationization process. Despite the immense potential of environmental sound classification, the accuracy and efficiency of automated classification often fall short of expectations. This paper proposes an environmental sound classification algorithm that integrates multiscale channel feature fusion, aiming to significantly reduce computational complexity while improving classification accuracy. The proposed algorithm comprises two modules: a multiscale channel feature fusion module and a selective feature fusion module, which dynamically merge temporal and frequency domain features. Finally, a series of ablation experiments are conducted and compared with other mainstream algorithms in environmental sound classification, revealing that the proposed algorithm possesses smaller parameter count and higher classification accuracy, thus substantiating the effectiveness of the multiscale channel feature fusion algorithm.

1 INTRODUCTION

Environmental sound refers to the non-linguistic sounds that surround us in our daily lives, providing important clues about our surroundings. Studying environmental sound is of significant importance for understanding the environment, predicting environmental changes, and improving the relationship between humans and the environment. However, the current performance of environmental sound classification is not ideal, and this field faces numerous challenges. As a non-stationary signal, environmental sound covers a wide range of audio sources, making it difficult to represent with a single template. Moreover, environmental sound itself is similar to background noise, further complicating the classification process. To address the main challenge of extracting effective audio features in environmental sound classification, this paper proposes an algorithm based on multiscale channel feature fusion and introduces a Multiscale Channel Feature Fusion Module (MFFM). By processing the mel spectrogram features through mel filters, the algorithm extracts both temporal and frequency domain features. It then utilizes a method of attention-based multi-channel feature fusion to generate a novel three-dimensional feature map, thereby further enhancing the accuracy of environmental sound classification.

2 RELATED WORK

Early research on environmental sound classification primarily focused on traditional classification methods (Temko, A.- Piczak, K.), with limited attention given to audio feature extraction, leading to lower classification accuracy. In recent years, the integration of deep neural networks with feature extraction has gradually become the mainstream approach for environmental sound classification, showcasing excellent generalization performance. For instance, Piczak proposed a method for environmental sound classification using a two-layer convolutional network, pioneering the application of convolutional neural networks in this field (Piczak, 2015). Dong, from both the temporal and frequency domains, analyzed audio signals and verified the effectiveness of processing audio data from both temporal and frequency perspectives (Zhang, 2018). Zhang et al. utilized a CRNN model based on

Zhao, W., Wang, H., Chen, Y., Pan, X., Zhang, K. and Bai, Z.
An Environmental Sound Classification Algorithm Based on Multiscale Channel Feature Fusion.
DOI: 10.5220/0012273800003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 56-60
ISBN: 978-989-758-677-4

attention mechanism to capture the correlation between the temporal and frequency domains, achieving high classification accuracy. However, the challenge of extracting effective representations for audio features still persists. Therefore, this paper proposes an algorithm model for environmental sound classification based on multiscale channel feature fusion, consisting of a temporal-frequency feature extraction module and a selective feature fusion module. Through ablative experiments on temporal and frequency features and comparative experiments on selective feature fusion, the effectiveness of the proposed algorithm in this paper is ultimately demonstrated.

3 MULTISCALE CHANNEL FEATURE FUSION

This chapter presents the algorithm for multiscale channel feature fusion. Since audio is typically composed of three physical attributes: frequency, time, and amplitude, and the human auditory system perceives sounds through neural encoding along these three dimensions, we propose a feature extraction module that mimics the human auditory system to capture the frequency-time characteristics of environmental sounds.

3.1 Module for Temporal and Spectral Feature Extraction

By simulating the mechanisms of the human auditory system, this study adopts a dual perspective from the time domain and frequency domain to mimic how the human ear classifies environmental sounds. By applying frequency attention and time attention mechanisms, we can dynamically select and fuse frequency domain and time domain features to obtain more accurate representations of environmental sound. As the spectrogram processed by mel filters is a spectro-temporal representation, the feature extraction module in this section separately extracts features from the time domain and frequency domain. The specific module structure is illustrated in Figure 1.

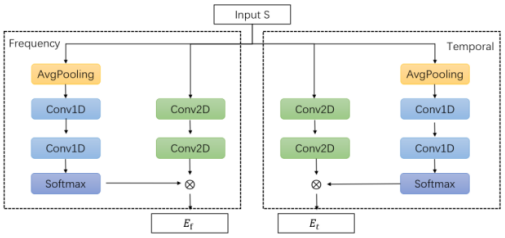


Figure 1: Module for Temporal and Spectral Feature Extraction.

3.2 Selective Feature Fusion Modules

Based on the temporal and spectral features obtained from Section A, we propose the Selective Feature Fusion Module (SFFM) for dynamically selecting and fusing temporal and spectral features. Figure 2 illustrates the detailed architecture of this module.

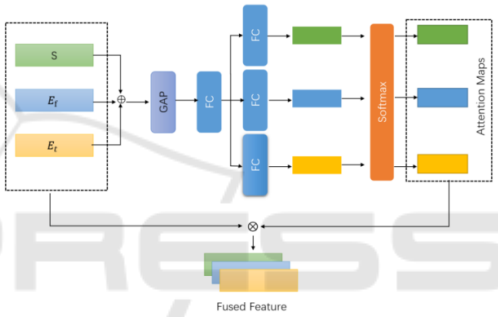


Figure 2: Selective Feature Fusion Module.

The input of SFFM comprises three components, including temporal feature map Et, frequency feature map Ef, and Log Mel spectrogram S. Initially, we merge the obtained S, Et, and Ef through element-wise addition to create a new feature mapping E. Subsequently, we perform global average pooling (GAP) to derive the global feature mapping gap.

$$gap = \frac{1}{F*T}\sum_{i\le F,\ j\le T} E_{ij} \quad (1)$$

After passing through a fully connected (FC) layer with a non-linear transformation, three additional FC layers are utilized to learn crucial features for each channel. A softmax layer is then applied to generate feature mappings. Subsequently, matrix multiplication is performed between the three inputs and feature mappings to obtain weighted feature maps. Following channel-wise summation, the fused three-dimensional temporal-frequency feature map E is obtained, which encapsulates abundant temporal and frequency information.

$$E = (S * softmax(fc(gap)) \oplus S_f * softmax(fc(gap)) \oplus S_t * softmax(fc(gap))) \quad (2)$$

4 EXPERIMENTAL SETUP AND RESULT ANALYSIS

4.1 Experimental Setup

In this part of the experiment, we choose the sampling rate of 11025 to balance the computational efficiency and obtain higher classification accuracy. The extracted logmel feature frames have a duration of 1024, use a 60 mel filter, and utilize 50% window overlap. The training process of the network model consists of 120 epochs with a batch size of 128. For ESC-10, ESC-50 and UrbanSound8K datasets, five and ten cross-validations were used, respectively. In addition, we choose to use classification cross entropy as a loss function and adopt L2 regularization to enhance the generalization ability of the model.

4.2 Temporal-Frequency Feature Ablation Experiment

To validate the effectiveness of the multi-scale channel feature fusion method, a series of ablation experiments were conducted using the UrbanSound8K, ESC-10, and ESC-50 datasets for evaluation. Table 1 presents the results of the temporal-frequency feature ablation experiments in this section. The experimental results demonstrate that, in the UrbanSound8K dataset, the classification accuracy fluctuates between 94% and 95% when using a single feature or fusing two features. However, when all three features are fused together, the classification accuracy reaches 96.7%. In the ESC-10 and ESC-50 datasets, the classification accuracy achieved after channel feature fusion is 80% for both datasets. Removing one or two features leads to a significant decrease in the classification accuracy. Therefore, the proposed channel feature fusion method has advantages in handling datasets with a smaller sample size and a larger number of categories.

Table 1: Ablation results of time-frequency characteristics.

Time-domain feature	Frequency-domain feature	Logmel feature	UrbanSound8K	ESC-10	ESC-50
√	×	×	95.1%	70.0%	74.0%
×	√	×	95.3%	72.5%	74.5%
×	×	√	94.5%	77.5%	75.0%
√	√	×	95.1%	72.5%	69.5%
×	√	√	95.3%	75.0%	68.0%
√	×	√	95.3%	72.5%	70.0%
√	√	√	96.7%	82.0%	80.0%

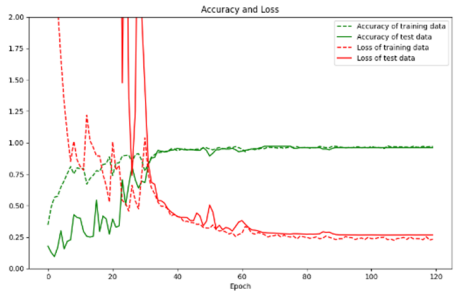
4.3 Select Feature Fusion Comparative Experiment

In the feature fusion part, the SFFM proposed in this paper is compared with the average superposition, weighted superposition, horizontal concatenation and vertical concatenation feature fusion methods on UrbanSound8K data set, so as to verify the feature representation ability of SFFM after feature fusion.

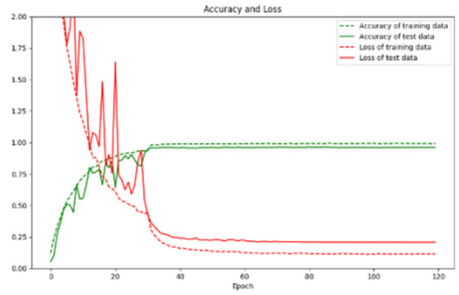
Table 2: Comparison between feature fusion and other fusion modes.

Fusion mode	Feature size after fusion	UrbanSound8K	ESC-10	ESC-50
Average stack	(60,44,1)	94.9%	75.0%	75.0%
Weighted stack	(60,44,1)	94.8%	77.5%	75.0%
Horizontal splicing	(120,44,1)	94.7%	75.0%	64.0%
Vertical splicing	(60,88,1)	96.3%	72.5%	80.0%
Selective feature fusion	(60,44,3)	96.7%	82.0%	80.0%

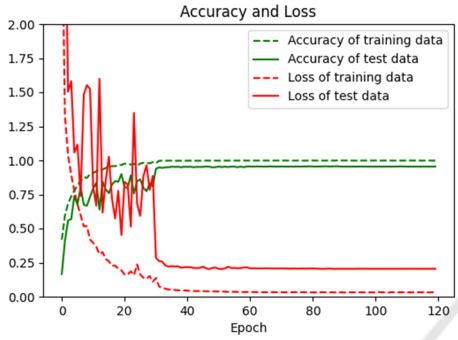
The comparative experimental results are shown in Table 2, and it can be clearly seen that the SFFM feature fusion method proposed in this paper shows the optimal classification effect on each data set.In order to better display the experimental results, this section visualizes the experimental classification results and presents them from multiple perspectives by using curves and confusion matrix.



(a)



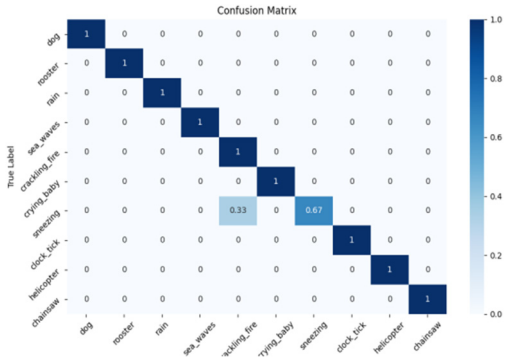
(b)



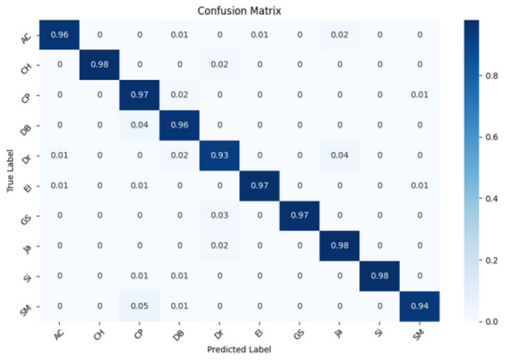
(c)

Figure 3: Accuracy curve and loss curve of ESC-10, ESC-50, UrbanSound8K.

The curves depicting the classification accuracy and loss of the environmental sound classification algorithm used in this study, for the ESC-10, ESC-50, and UrbanSound8K datasets, are presented in Figure 3. Specifically, Figure (a) corresponds to the ESC-10 dataset, Figure (b) represents the ESC-50 dataset, and Figure (c) relates to the UrbanSound8K dataset. In these curves, the solid red line represents the loss curve of the test set, while the dashed red line represents the loss curve of the training set. On the other hand, the solid green line depicts the accuracy curve of the test set, while the dashed green line signifies the accuracy curve of the training set.



1)



2)

Figure 4: ESC-10 Confusion matrixAND and UrbanSound8K confusion matrix.

4.4 Comparison with SOTA Methods at Home and Abroad

Table 3 compares the environmental sound classification method proposed in this paper based on multi-scale channel feature fusion with other mainstream methods at home and abroad. On the UrbanSound8K dataset, our method achieves 97.3% classification accuracy, which is 2% higher than the enhanced DCNN model. It is slightly lower than Mushtaq's NAA method and Inik's combined CNN and PSO method. However, both methods achieve high accuracy while significantly increasing computational complexity and parameter size. In contrast, the network architecture adopted in this paper is a stacked four-layer neural network with a parameter size of only 0.44M, which is smaller than all current mainstream methods and saves computing resources.

Table 3: Compares SOTA mainstream methods at home and abroad on different datasets.

Year	Author	Method	Parameter quantity	ESC-10	ESC-50	UrbanSound 8K
2015	Piczak	Piczak-CNN	26M	73.00%	64.50%	73.70%
2016	Dai et al.	M18	8.7M	/	/	71.68%
2018	Zhang et al.	Improved VGG8	/	91.70%	83.90%	83.70%
2019	Su et al.	CNN6	6.6M	/	85.60%	93.40%
2020	Mushtaq et al.	DCNN	3.0M	94.90%	89.20%	95.30%
2021	Zhang et al.	ACRNN	3.81M	93.70%	86.10%	/
2021	Mushtaq et al.	NAA	112M	99.04%	97.57%	99.49%
2021	S. Luz et al.	Handcrafted+Deep	1.48M	/	86.20%	96.80%
2023	Inik	CNN+PSO	45.9M	98.64%	96.77%	98.45%
This Paper		SFFM	0.44M	96.11%	95.90%	97.30%

5 CONCLUSION

In this paper, we propose a feature optimization method called multi-scale channel feature fusion. It involves extracting logmel features in both the time and frequency domains for sound classification. Subsequently, an attention mechanism is employed to fuse the original features, time-domain features, and frequency-domain features across channels, enabling efficient classification of environmental sounds. By comparing our method with current state-of-the-art approaches in terms of accuracy and parameter size, we provide a comprehensive evaluation of the advantages and disadvantages of our proposed method.

ACKNOWLEDGMENTS

This work was financially supported by Major scientific and technological innovation Project of Shandong Key R & D Plan "Smart and Healthy Air Industry Project based on New Generation Information Technology" , and Key R & D projects of Shandong Province (2020JMRH0201).

REFERENCES

Temko, A., Nadeu, C. Acoustic Event Detection in Meeting-Room Environments[J]. *Pattern Recognition Letters* 2009, 30 (14), 1281–1288. https://doi.org/10.1016/j.patrec.2009.06.009.

Gupta, S.; Karanath, A.; Mahrifa, K.; Dileep, A.; Thenkanidiyoor, V. Segment-Level Probabilistic Sequence Kernel and Segment-Level Pyramid Match Kernel Based Extreme Learning Machine for Classification of Varying Length Patterns of Speech[J]. *International Journal of Speech Technology* 2019, 22 (1), 231–249. https://doi.org/10.1007/s10772-018-09587-1.

Stowell, D.; Giannoulis, D.; Benetos, E.; Lagrange, M.; Plumbley, M. Detection and Classification of Acoustic Scenes and Events[J]. *IEEE Transactions On Multimedia* 2015, 17 (10), 1733–1746. https://doi.org/10.1109/TMM.2015.2428998.

Piczak, K.; ACM. ESC: Dataset for Environmental Sound Classification[C]; 2015; pp 1015–1018. https://doi.org/ 10.1145/2733373.2806390.

Piczak KJ. Environmental Sound Classification With Convolutional Neural Networks.[C] In: Erdogmus D, Akcakaya M, Kozat S, Larsen J, eds. 2015 IEEE International Workshop on Machine Learning for Signal Processing. *IEEE International Workshop on Machine Learning for Signal Processing. IEEE Signal Processing Soc*; Northeast Univ; Intel; 2015.

Zhang, Z.; Xu, S.; Cao, S.; Zhang, S. Deep Convolutional Neural Network with Mixup for Environmental Sound Classification[J]; *Pattern Recognition and Computer Vision*, PT II, Eds.; 2018; Vol. 11257, pp 356–367. https://doi.org/10.1007/978-3-030-03335-4_31.

Dai W, Dai C, Qu S, et al. Very Deep Convolutional Neural Networks for Raw Waveforms[C], *IEEE International Conference on Acoustics*, 2017,421-425.

Su, Y.; Zhang, K.; Wang, J.; Zhou, D.; Madani, K. Performance Analysis of Multiple Aggregated Acoustic Features for Environment Sound Classification[J]. *Applied Acoustics* 2020, 158. https://doi.org/10.1016/j.apacoust.2019.107050.

Mushtaq, Z.; Su, S. Environmental Sound Classification Using a Regularized Deep Convolutional Neural Network with Data Augmentation[J]. *Applied Acoustics* 2020, 167. https://doi.org/10.1016/j.apacoust.2020.107389.

Zhang, Z.; Xu, S.; Zhang, S.; Qiao, T.; Cao, S. Attention Based Convolutional Recurrent Neural Network for Environmental Sound Classification[J]. *Neurocomputing* 2021, 453, 896–903. https://doi.org/10.1016/j.neucom.2020.08.069.

Mushtaq, Z.; Su, S.; Tran, Q. Spectral Images Based Environmental Sound Classification Using CNN with Meaningful Data Augmentation[J]. *Applied Acoustics* 2021, 172. https://doi.org/10.1016/j.apacoust.2020.107581.

Luz, J.; Oliveira, M.; Araujo, F.; Magalhaes, D. Ensemble of Handcrafted and Deep Features for Urban Sound Classification[J]. *Applied Acoustics* 2021, 175. https://doi.org/10.1016/j.apacoust.2020.107819.

Inik, O. CNN Hyper-Parameter Optimization for Environmental Sound Classification[J]. *Applied Acoustics* 2023, 202. https://doi.org/10.1016/j.apacoust.2022.109168.