sensors

MDPI

Article

High Accurate Environmental Sound Classification: Sub-Spectrogram Segmentation versus Temporal-Frequency Attention Mechanism

Tianhao Qiao, Shunqing Zhang *, Shan Cao and Shugong Xu

Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China; qiaotianhao@shu.edu.cn (T.Q.); cshan@shu.edu.cn (S.C.); shugong@shu.edu.cn (S.X.)
* Correspondence: shunqing@shu.edu.cn

Abstract: In the important and challenging field of environmental sound classification (ESC), a crucial and even decisive factor is the feature representation ability, which can directly affect the accuracy of classification. Therefore, the classification performance often depends to a large extent on whether the effective representative features can be extracted from the environmental sound. In this paper, we firstly propose a sub-spectrogram segmentation with score level fusion based ESC classification framework, and we adopt the proposed convolutional recurrent neural network (CRNN) for improving the classification accuracy. By evaluating numerous truncation schemes, we numerically figure out the optimal number of sub-spectrograms and the corresponding band ranges, and, on this basis, we propose a joint attention mechanism with temporal and frequency attention mechanisms and use the global attention mechanism when generating the attention map. Finally, the numerical results show that the two frameworks we proposed can achieve 82.1% and 86.4% classification accuracy on the public environmental sound dataset ESC-50, respectively, which is equivalent to more than 13.5% improvement over the traditional baseline scheme.

Keywords: environmental sound classification; convolutional recurrent neural network; sub-spectrogram segmentation; score level fusion; temporal-frequency attention mechanism

check for updates

Citation: Qiao, T.; Zhang, S.; Cao, S.; Xu, S. High Accurate Environmental Sound Classification: Sub-Spectrogram Segmentation versus Temporal-Frequency Attention Mechanism. *Sensors* **2021**, *21*, 5500. https://doi.org/10.3390/s21165500

Academic Editors: Leon Rothkrantz and Ahmet Kondoz

Received: 23 April 2021
Accepted: 10 August 2021
Published: 16 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As a key technology for recognizing and analyzing environmental audio signals, environmental sound classification (ESC) [1] is rapidly developed during the past few years with broad applications in home automation, machine hearing, as well as surveillance. Compared with traditional sound classification tasks, such as music or speech recognition [2], the development of this field is relatively slow. This is due to the reason that ESC requires to cover a wide range of frequency spectrum, non-stationary characteristic, and noise-like fluctuations [3–5].

Since traditional ESC methods usually consist of feature extraction and feature-based classification processes, a general extension scheme is to improve the classification accuracy in a block-by-block manner. For example, zero crossing rate, audio tone, and short-time energy have been proposed in [6] to improve the feature extraction capability in the low noise environments, while a significant computational complexity is usually required to achieve a reasonable classification accuracy. To solve this problem, extracting features in frequency domain and using temporal-frequency spectrograms to represent environmental sounds becomes the most widely used method recently [7] on the basis of making full use of the recent developments in the field of image processing tasks. Mel-frequency cepstrum coefficient (MFCC) [8] could be one example, while log mel spectrogram (Logmel) [9] and log Gammatone spectrogram [10] are more popular in the recent years. Another approach is to update the feature-based classification block, and typical examples include K-nearest neighbors [11], random forest [12], support vector machine [13], and Gaussian mixture

model [14]. With the recent development of supervised learning technologies, the feature-based classification algorithms have also been extended to cover dictionary learning [4], matrix factorization [15], and deep learning based solutions, such as deep neural networks (DNN) [16].

In recent years, the record of achievable classification accuracy has been updated simply through different combinations of feature extraction methods and feature-based classification blocks [9,17–19]. For example, when MFCC is combined with two different DNN structures, multi-layer perception (MLP) and convolutional neural network (CNN), the classification accuracy of 44.9% and 53.1% can be achieved in the public environmental sound dataset ESC-50, and when log Gammatone is used in combination with CNN, the classification accuracy can reach 78.9% in the same dataset. However, the above schemes failed to incorporate some specific domain knowledge and the achievable classification accuracy is in general limited. In addition, we should note that most of the existing researches focus on the improvement of neural networks, which is generally applicable to any machine learning field, but few consider the improved processing method of features specific to environmental sound. As far as we are aware, the following issues need to be addressed at the current stage.

- *Sub-spectrogram segmentation:* It is very necessary to study the spectrograms of environmental sound more carefully. This is because the low frequency spectrum usually contains more fruitful information, as explained in [10]. Although a straight-forward sub-spectrogram segmentation as proposed in [20] is shown to be effective to improve the acoustic scene classification accuracy, the extension to ESC tasks still remains open. In addition, according to the existing literature, the number of sub-spectrogram segments, as well as the truncation rules need to be optimized as well;
- *Attention mechanism:* Another possible approach to improve the ESC performance is to incorporate the attention mechanism like human beings [21–25] in the convolutional feature layers, either through the temporal [24], frequency [26], channel [27] domain information, or even hybrid of them [27]. However, the previous joint attention scheme [27] focuses on combining the temporal and channel knowledge without considering the frequency domain characteristics, and the joint time-frequency feature is not fully exploited. As shown later, with the joint time and frequency domain attention, the ESC accuracy can be greatly improved;
- *Recurrent architecture with data augmentation:* The sound of many consecutive frames, such as *helicopter*, has strong correlations in the time domain, and the prediction via recurrent architecture will be possible. As shown in [28], exploiting the correlations among different scales of sequences can be applied to improve the classification accuracy as well. However, this method usually requires a large amount of data to support and comes with the problem of limited dataset. Therefore, it is very necessary to jointly consider the effective methods for expanding the dataset, such as mixup [29] and SpecAugment [30].

In this paper, a sub-spectrogram segmentation [31] (Part of this paper has been published in 2019 IEEE International Workshop on Signal Processing Systems) mechanism has been firstly proposed to address the above concerns, which truncates the entire spectrogram into different pieces in order to conduct experiments separately. Score level fusion has been adopted to combine different classification results from different sub-spectrograms. By evaluating numerous truncation schemes, we numerically figure out the optimal number of sub-spectrograms and the corresponding band ranges. Based on that, we propose a joint attention mechanism with temporal and frequency domain to adjust the temporal-frequency feature map, which can be similarly regarded as automatically assigning the weight map to the feature map. Numerical results show that the two frameworks we proposed can achieve 82.1% and 86.4% classification accuracy on ESC-50 dataset, respectively, which is equivalent to more than 13.5% improvement over the traditional baseline scheme.

The rest of this paper contains the following parts. We first gave a brief introduction to log Gammatone spectrogram and different types of DNN in Section 2, and the overviews of

our proposed methods are introduced in Section 3. The proposed sub-spectrogram segmentation and temporal-frequency attention mechanism based ESC classification frameworks are, respectively, introduced in Sections 4 and 5. In addition, we gave a demonstration of the numerical experiments in Sections 6 and 7 gives a summary of this paper at the end.

2. Preliminary

In this section, we give a brief introduction to the well-known log Gammatone spectrogram and different types of DNNs.

2.1. Log Gammatone Spectrogram

By performing T point discrete short-time Fourier transform (STFT) on the sampled time domain audio signal $s(t)$, the energy spectrum density, $|S(m,n)|^2$, can be obtained, and the formula is as follows,

$$|S(m,n)|^2 = \left| \sum_{t=\frac{nT}{2}+1}^{\frac{(n+2)T}{2}} s(t) \cdot e^{-\frac{j2\pi mt}{T}} \right|^2, \tag{1}$$

for $m \in [1, T/2]$, $n \in [0, N-1]$. Then apply K order Gammatone-filter banks to it to determine the log Gammatone spectrogram, and the mathematical expression of this process can be expressed as,

$$S_{GT}(n,k) = \log\left(\sum_{m=1}^{T/2} |S(m,n)|^2 \cdot H(m,k) \right), \tag{2}$$

for $k \in [1, K]$, where $H(m,k)$ denotes the frequency response of the kth Gammatone-filter in the mth sub-band. In addition, $H(m,k)$ and the associated time domain impulse response, $h(t)$, can be, respectively, given by,

$$\begin{aligned} H(m,k) &= \frac{C}{2}(k-1)!(2\pi b)^{-k}[P(m) + P^*(-m)], \\ P(m) &= e^{i\phi}[1 + i\frac{(m-f_0)}{b}]^{-k}, \end{aligned} \tag{3}$$

$$h(t) = C \cdot t^{k-1} e^{-2\pi bt} cos(2\pi f_0 t + \phi), \forall\, t > 0, \tag{4}$$

where f_0 and ϕ denote the center frequency and the corresponding phase information, k and $b > 0$ denote the order of filter and the decay rate, and C is an empirical constant to adjust the entire value.

In the actual system, usually as shown in [10], select $\{H(m,k)\}$ to model the characteristics over the entire frequency band, (f_L, f_H), e.g., from zero to half of the sampling frequency. Moreover, according to [27], the log Gammatone spectrogram is often a two-dimensional channel tensor, which contains itself and its delta information, e.g., $\mathbf{S}_{GT}(n,k)$, and Figure 1 shows the log Gammatone spectrograms of four typical sound classes.

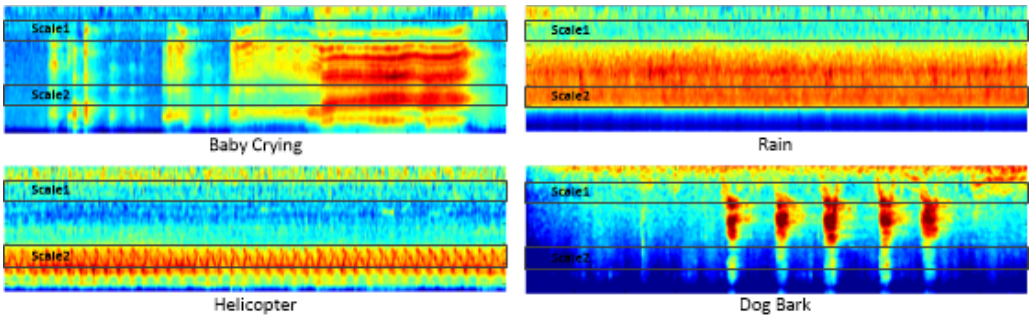


Figure 1. Four typical log Gammatone spectrograms of crying baby, rain, helicopter, and dog bark over the entire frequency band, where the horizontal axis represents time dimension and the vertical axis represents frequency dimension.

2.2. Deep Neural Networks

In general, DNN refers to a more powerful neural network formed by connecting multiple layers of neurons, such as multi-layer perception (MLP), convolutional neural network (CNN), and recurrent neural network (RNN). The design philosophy of MLP and CNN is roughly the same, and the difference is that the neurons within each layer are isolated and neurons across neighboring layers are fully connected in MLP, while the neurons across neighboring layers are connected together through convolution kernels and pooling operations in CNN. In addition, CNN can learn local patterns among different input elements with the help of convolutional architecture, for instance, image pixels or environmental sound spectrograms. RNN is proposed to solve the problem of the temporal correlation among different input vectors or patterns, which is not considered in the above two structures. It cannot only use previous frame-level functions, but also learn complex temporal dynamics. In previous researches, DNN has been proven to be able to handle many challenging tasks in the fields of ESC and computer vision by combining different architectures together.

3. Overview of the Proposed High Accurate ESC

In this part, we put forward two approaches for environmental sound classification, namely *sub-spectrogram segmentation* and *temporal-frequency attention*, where an overview of them is shown in Figure 2.

3.1. Overview

In general, the ESC task relies on the observed sound signal $s(t)$ or the equivalent energy spectrum $|S(m,n)|^2$ to classify different sound classes. The mathematical expression of the classification task of N_{cls} classes of sounds is,

$$\mathbf{p}_{N_{cls}} = F\Big(\Big\{|S(m,n)|^2\Big\}\Big), \tag{5}$$

where $\mathbf{p}_{N_{cls}} = [p_1, p_2, \ldots, p_{N_{cls}}]^T$ denotes the probability distribution across N_{cls} sound classes. The non-linear function $F(\cdot)$ is directly approximated by its equivalent log Gammatone spectrogram and the corresponding neural network defined by θ in the traditional approaches, e.g., $\mathbf{p}_{N_{cls}} = G(\{\mathbf{S}_{GT}(n,k)\};\theta)$.

3.2. Sub-Spectrogram Segmentation

We truncated the whole spectrograms into N_{ss} parts, e.g., $(f_L, f_1), \ldots, (f_{N_{ss}-1}, f_H)$ instead of generating the log Gammatone spectrogram based on the entire frequency band, and use score level fusion when performing the decision. The overall operations can be described mathematically as follows,

$$\mathbf{p}_{N_{cls}} = \sum_{i=1}^{N_{ss}} \omega^i \mathbf{p}^i_{N_{cls}} = \sum_{i=1}^{N_{ss}} \omega^i G\Big(\Big\{\mathbf{S}^i_{GT}(n,k)\Big\};\theta\Big), \tag{6}$$

where $\mathbf{p}^i_{N_{cls}}$ and ω^i denote the score of the ith sub-spectrogram and the fusion weight, respectively, and $\sum_{i=1}^{N_{ss}} \omega^i = 1$. $\{\mathbf{S}^i_{GT}(n,k)\}$ defines the generated log Gammatone spectrograms based on the ith band (For illustration purpose, we define $f_0 = f_L$ and $f_{N_{ss}} = f_H$.), e.g., from f_{i-1} to f_i, and $G(\cdot;\theta)$ represents a non-linear mapping between log Gammatone spectrogram and classification results.

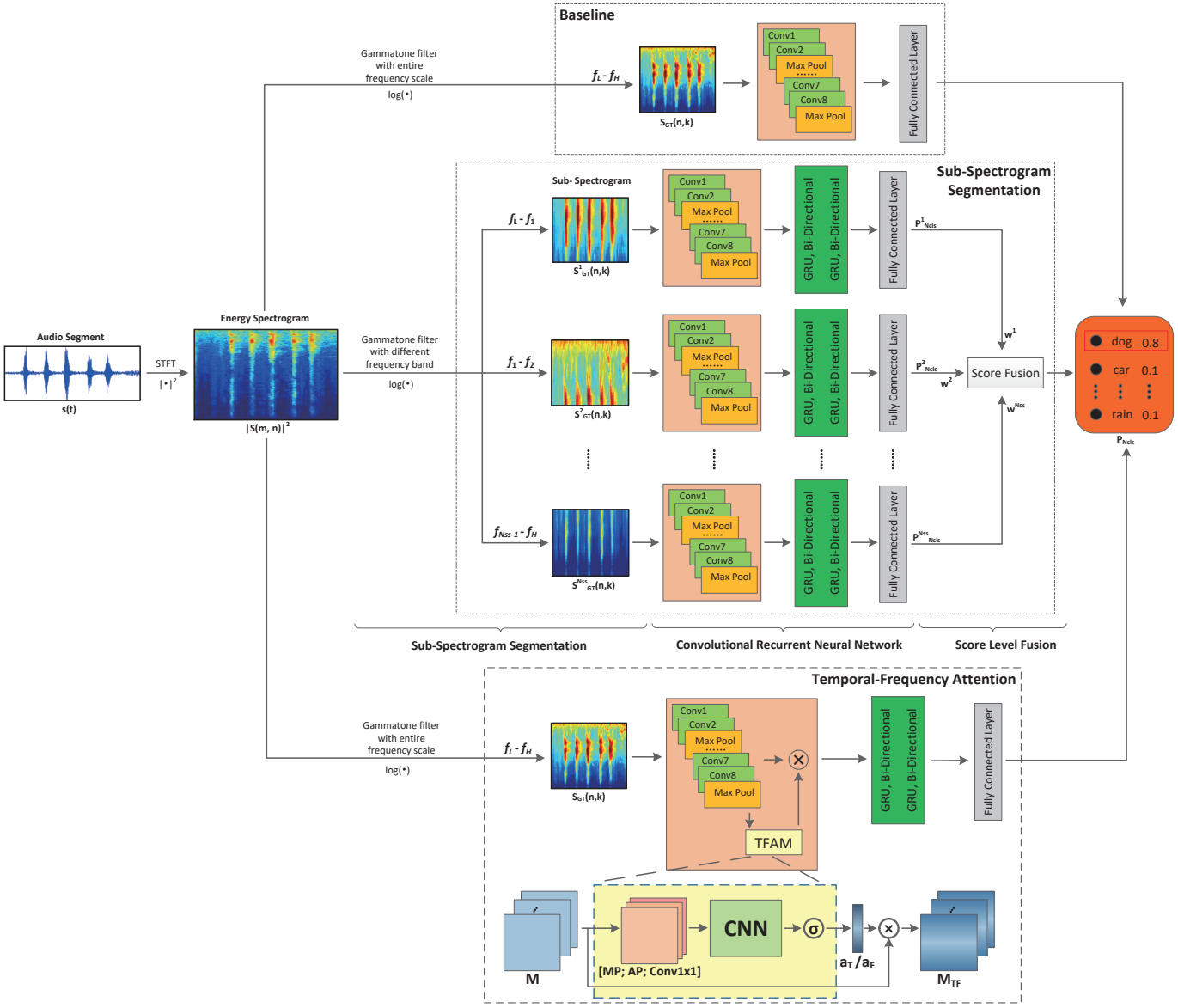


Figure 2. Compare the baseline system, the proposed sub-spectrogram segmentation system and the proposed temporal-frequency attention mechanism system. In this figure, the first branch denotes the baseline system, which extracts log Gammatone spectrogram features on the entire frequency band, the second branch denotes the proposed sub-spectrogram segmentation method, which extracts log Gammatone spectrogram on several sub-frequency bands as illustrated, and the last branch denotes the proposed temporal-frequency attention mechanism (TFAM) and temporal-frequency attention system.

3.3. Temporal-Frequency Attention

The above sub-spectrogram segmentation mechanism only considers the frequency domain, which ignores the temporal domain characteristics. To address this issue, we propose a temporal-frequency attention mechanism (TFAM) as illustrated in Figure 2. Given the input log Gammatone spectrogram, $\{\mathbf{S}_{GT}(n,k)\}$, we first use CNN to extract temporal-

frequency representations. Mathematically, we generate the feature maps, $\mathbf{M} \in \mathbb{R}^{T \times F \times C}$, on top of the log Gammatone spectrograms according to the following expression,

$$\mathbf{M} = H_1(\{\mathbf{S}_{GT}(n,k)\};\theta_1), \tag{7}$$

where T, F and C represents the dimension of feature maps and $H_1(\cdot;\theta_1)$ represents the non-linear transformation provided by CNN.

In order to keep the implementation complexity, we restrict the attention map according to $\mathbf{W}_{AT} \otimes \mathbf{1}$, where $\mathbf{W_{AT}} \in \mathbb{R}^{T \times F}$ denotes the temporal-frequency attention patterns, $\mathbf{1} \in \mathbb{R}^{C}$ denotes an all-one vector with dimension C, and $\otimes$ is Kronecker product as defined in [32]. With the generated attention map, the overall operations can be described as follows,

$$\mathbf{p}_{N_{cls}} = G(\mathbf{M} \cdot (\mathbf{W}_{AT} \otimes \mathbf{1});\theta), \tag{8}$$

where $\cdot$ denotes the element-wise multiplication, and $G(\cdot;\theta)$ represents a non-linear mapping between log Gammatone spectrogram and classification results.

4. Proposed Sub-Spectrogram Segmentation Based Classification Framework

In this section, we respectively introduce sub-spectrogram segmentation based feature extraction, CRNN based classification, and score level fusion, which are all components of the proposed sub-spectrogram segmentation based classification framework.

4.1. Sub-Spectrogram Segmentation

From Figure 1, we can see that the difference of behaviour in different scales of the spectrogram is really significant. Firstly, we divide the entire log Gammatone spectrogram into two parts, and the parameter settings are $N_{ss} = 2$, $f_L = 0$ kHz, $f_1 = 10$ kHz, and $f_H = 22.05$ kHz. The classification accuracy on ESC-50 dataset [33] changes with ω^1 as shown in Figure 3, and it can be concluded that as long as there is an appropriate weight assignment, the proposed sub-spectrogram segmentation can outperform the baseline system.

Secondly, we identify the optimal number of sub-spectrogram segments through experiments, e.g., $N_{ss}^{\star}$, and we provide extensive numerical studies. Specifically speaking, we evaluate the system performance under different values of N_{ss} and $\{f_i\}$, and Table 1 lists the results when using the optimal weight coefficients, $\{\omega^i\}$. From Table 1 we can see that the accuracy does *NOT* increase monotonically with regard to N_{ss}, and the optimal number of sub-spectrogram segments is $N_{ss}^{\star} = 4$.

Table 1. Classification accuracy under different values of N_{ss}.

N_{ss}	f_L (kHz)	$\{f_i\}$ (kHz)	f_H (kHz)	Accuracy
1	0	-	22.05	77.9%
2	0	10	22.05	79.9%
3	0	6, 10	22.05	81.7%
4	0	3, 6, 10	22.05	82.1%
5	0	3, 6, 10, 15	22.05	81.8%
6	0	3, 6, 10, 13, 16	22.05	81.3%

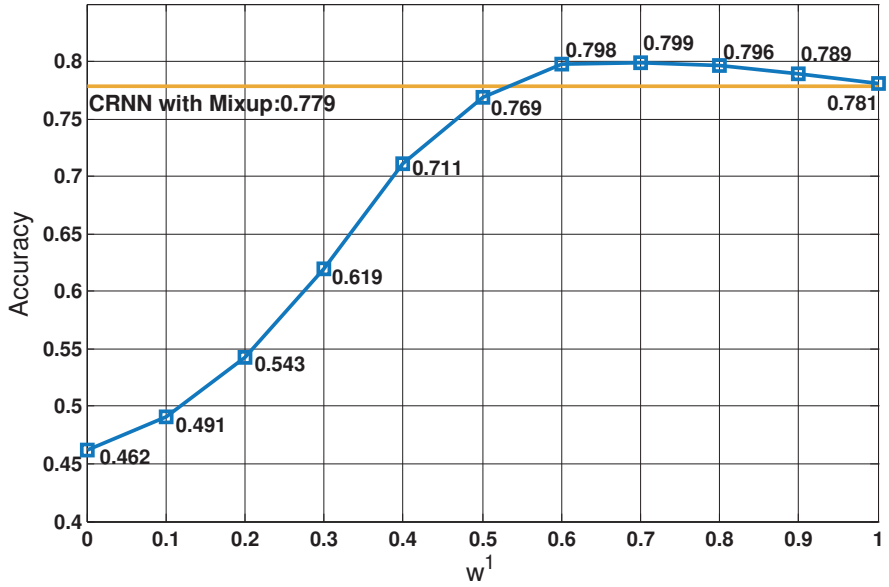


Figure 3. Classification accuracy with different weights. The blue line indicates the classification accuracy of different weights cases, while the orange line indicates the accuracy of CRNN with mixup system.

4.2. CRNN with Mixup

Inspired by the complementary modeling capabilities of CNN and RNN, we combine them into a unified architecture called *convolutional recurrent neural network (CRNN)*, which can be represented as the approximate original non-linear function $G(\cdot;\theta)$. The complementary modeling function mentioned here, respectively, refers to using a convolution kernel with a small receptive field on spectrogram features to capture the local spectro-temporal pattern and learning the temporal relationship of the environmental sound features. Specifically, in this system, the learned features obtained through conventional convolutional layer will first be forwarded into the bi-directional gated recurrent unit (GRU) for temporal processing, and then the score of the ith sub-spectrogram, $\mathbf{p}^i_{N_{cls}}$, can be obtained. In addition, detailed architecture of the proposed CRNN and its parameters are presented in Table 2.

Table 2. Architecture of the proposed convolutional recurrent neural network (CRNN).

Layer	Nums of Filters	Filter Size	Stride	Output Size
Conv1	32	(3, 3)	(1,1)	(128, 128, 32)
Conv2	32	(3, 3)	(1,1)	(128, 128, 32)
Pool1	-	-	(4, 2)	(32,64,32)
Conv3	64	(3, 1)	(1, 1)	(32, 64, 64)
Conv4	64	(3, 1)	(1, 1)	(32, 64, 64)
Pool2	-	-	(2, 1)	(16,64,64)
Conv5	128	(1, 3)	(1, 1)	(16, 64, 128)
Conv6	128	(1, 3)	(1, 1)	(16, 64, 128)
Pool3	-	-	(1, 2)	(16,32,128)
Conv7	256	(3, 3)	(1, 1)	(16, 32, 256)
Conv8	256	(3, 3)	(1, 1)	(16, 32, 256)
Pool4	-	-	(2, 2)	(8, 16, 256)
GRU1	256	-	-	(16, 256)
GRU2	256	-	-	(16, 256)
FC1	nums of classes	-	-	(nums of classes)

For avoiding overfitting that may be caused by the limited training dataset, here we use the data augmentation method, mixup, for constructing virtual training data to achieve the purpose of expanding the training distribution [34]. mixup generates the virtual training data by mixing two training samples, e.g., to attain a mixed virtual feature by mixing a *crying baby* log Gammatone spectrogram and a *dog bark* log Gammatone spectrogram, the formula can be expressed as

$$\{\widetilde{S}^i_{GT}(n,k)\} = \lambda\{S^i_{GT}(n,k)\}_j + (1-\lambda)\{S^i_{GT}(n,k)\}_{j'}, \tag{9}$$

where $\{S^i_{GT}(n,k)\}_j$ and $\{S^i_{GT}(n,k)\}_{j'}$ are two randomly selected samples in log Gammatone spectrograms of training data. Correspondingly, labels should also be mixed in the same ratio. λ is decided by two hyper-parameters, α and $\lambda \sim \text{Beta}(\alpha, \alpha)$ [34].

Table 3 shows that using CRNN or mixup can increase the classification accuracy by 2.3% and 3.2%, respectively, while CRNN can increase the classification accuracy by 5% higher than the baseline system.

Table 3. Classification accuracy of whether to use CRNN or mixup.

Network	Mixup	Accuracy
CNN	×	72.9%
CRNN	×	75.2%
CNN	√	76.1%
CRNN	√	77.9%

4.3. Score Level Fusion

Finally, we experimented to identify the optimal weights, $\{\omega^{\star,i}\}$, in score level fusion. We can obtain the optimal weights by exhaustively searching over all possible combinations of $\{\omega^i\}$, and the classification accuracy are shown in Table 4. According to the results, it can be seen that the accuracy of score level fusion can be improved by 2.3% to 3.9% over the uniform weights assignment.

Table 4. Classification accuracy under different score level fusion strategies.

N_{ss}	f_L (kHz)	$\{f_i\}$ (kHz)	f_H (kHz)	Fusion	Accuracy
2	0	10	22.05	×	76.2%
2	0	10	22.05	√	79.9%
3	0	6, 10	22.05	×	78.1%
3	0	6, 10	22.05	√	81.7%
4	0	3, 6, 10	22.05	×	79.8%
4	0	3, 6, 10	22.05	√	82.1%
5	0	3, 6, 10, 15	22.05	×	77.9%
5	0	3, 6, 10, 15	22.05	√	81.8%

5. Proposed Temporal-Frequency Attention Based Classification Framework

In the above method, segmentation boundaries and number of segments need to be optimized over a multi-dimensional search spaces, which is in general computationally prohibitive. In this section, we propose a low-complexity joint temporal-frequency domain searching mechanism to generate the temporal-frequency attention map and figure out a temporal-frequency attention based classification framework with data augmentation. The network structure use here is the same as the CRNN mentioned in Section 4.2.

It is worth mentioning that a similar attention method was introduced in [35]. Apart from the concatenation pattern of the temporal attention and frequency attention, the main difference is the method of obtaining the temporal and frequency attention map. In this paper, we first use the combination of 1×1 convolution and pooling to squeeze channel information, while only 1×1 convolution is used in [35]. The results in Table 5 are shown

that the combination way performs better. Then we use 3×3 convolution to learn attention map based on the channel-squeezed feature, while a global average pooling is used in [35]. The learnable attention network usually has ability to learn more valuable information from input feature.

Table 5. Classification accuracy comparison for max and average pooling, and 1×1 convolution.

Network	Max Pooling	Avg Pooling	1×1 Conv	Accuracy
CRNN	√	×	×	82.0%
CRNN	×	√	×	81.8%
CRNN	×	×	√	82.1%
CRNN	√	√	√	82.3%

5.1. Attention Map Generation

In order to efficiently search the most important temporal-frequency features of an audio spectrogram, we propose a temporal-frequency attention mechanism (TFAM) as shown in Figure 2. Different from the previous sub-spectrogram segmentation based scheme, TFAM directly focuses on the most important frames and frequency bands through training samples, which is more or less the same as semantic segmentation in computer vision tasks. By applying TFAM, the most important temporal-frequency blocks are automatically cached and selected by multiplying an attention map, which eventually helps to the classification tasks thereafter.

To generate the attention map $\mathbf{W_{AT}}$ in (8), we have,

$$\mathbf{W_{AT}} = H_2(\mathbf{g}(t,f);\theta_2), \tag{10}$$

where $H_2(\cdot;\theta_2)$ represents the non-linear transformation defined by CNN, and $\mathbf{g}(t,f)$ is the concatenated spatial map. Mathematically, $\mathbf{g}(t,f)$ can be obtained by the following expression.

$$\mathbf{g}(t,f) = \begin{cases} \max_c(\mathbf{M}(t,f,c)) \oplus & \text{max pooling} \\ \frac{1}{C}\sum_{c=0}^{C-1}\mathbf{M}(t,f,c) \oplus & \text{average pooling} \\ *_{1\times1}(\mathbf{M}(t,f,:)) & 1 \times 1 \text{ convolution} \end{cases}, \tag{11}$$

where $\oplus$ denotes the concatenation operation along the channel axis and $*_{1\times1}$ denotes the 1×1 convolution operation. Although we can apply different combinations of pooling and convolution operations, the numerical results in Table 5 show that the concatenated approach achieves better performance in terms of the classification accuracy.

Since the frequency domain characteristics of spectrogram features remains static over different time frames, we choose to process temporal and frequency domains separately as proposed in [36] instead of jointly processing them together as an image. Through this approach, we extract global temporal and frequency attention vectors, e.g., $\mathbf{a}_T \in \mathbb{R}^{T\times1\times1}$ and $\mathbf{a}_F \in \mathbb{R}^{1\times F\times1}$, and generate the final attention map $\mathbf{W_{AT}}$, according to

$$\mathbf{W_{AT}} = \mathbf{a}_T \otimes \mathbf{a}_F. \tag{12}$$

To obtain $\mathbf{a}_T$ and $\mathbf{a}_F$, we forward $\mathbf{g}(t,f)$ into a standard CNN network, which consists of three two-dimensional convolutions with 3×3 receptive field for learning the hidden representations and three one-dimensional max pooling layers for reducing the time, frequency or channel dimension. We can use the following formula to describe this process,

$$\mathbf{a}_T = \sigma(H_3(\mathbf{g}(t,f);\theta_3)), \tag{13}$$

$$\mathbf{a}_F = \sigma(H_4(\mathbf{g}(t,f);\theta_4)), \tag{14}$$

where $H_3(\cdot;\theta_3)$ and $H_4(\cdot;\theta_4)$ represent the non-linear transformations obtained by CNN, and $\sigma(\cdot)$ denotes the sigmoid activation function, which is use to restrict the vector elements to a range of (0, 1).

In order to further improve ESC accuracy, we cascade the proposed TFAM blocks after different CRNN pooling layers in Table 2, and the simulation results are shown in Table 6. A classification accuracy up to 83.1% can be reached if the proposed TFAM blocks are cascaded after each CRNN pooling layers, which also outperforms the previous sub-spectrogram segmentation mechanism.

Table 6. Classification accuracy comparison for different TFAM allocations.

Network	Layer Pool1	Layer Pool2	Layer Pool3	Layer Pool4	Accuracy
CRNN	√	×	×	×	82.3%
CRNN	×	√	×	×	82.0%
CRNN	×	×	√	×	81.9%
CRNN	×	×	×	√	81.9%
CRNN	√	√	×	×	82.0%
CRNN	√	×	√	×	82.6%
CRNN	√	×	×	√	82.0%
CRNN	×	√	√	×	81.9%
CRNN	×	√	×	√	82.6%
CRNN	×	×	√	√	82.1%
CRNN	√	√	√	×	82.2%
CRNN	√	√	×	√	82.0%
CRNN	√	×	√	√	81.6%
CRNN	×	√	√	√	82.7%
CRNN	√	√	√	√	83.1%

5.2. Data Augmentation Schemes

The entire network architecture with CRNN and four TFAM blocks are depicted in Figure 4, where the final learned feature map, $\mathbf{M} \cdot (\mathbf{W}_{AT} \otimes \mathbf{1})$, is forwarded to bi-directional GRU for the temporal processing. The overall classification results, $\mathbf{p}_{N_{cls}}$, are obtained from a fully connected network, with dimension 50×1.

Since we usually have limited sizes of datasets for environmental sound classification, SpecAugment [37] and mixup [34] strategies are adopted to increase the diversity of training sample. SpecAugment applies multiple temporal and frequency masking schemes to generate multiple masked log Gammatone spectrogram, and mixup adopts a randomly mixing strategy between training samples to generate virtual mixed log Gammatone spectrogram and extend the training distribution. By jointly utilizing the above data augmentation schemes, we have the classification accuracy results as listed in Table 7.

Table 7. Classification accuracy of whether to use mixup or SpecAugment.

Network	Mixup	SpecAugment	TFAM	Accuracy
CRNN	×	×	×	75.2%
CRNN	√	×	√	83.1%
CRNN	×	√	√	82.7%
CRNN	√	√	√	86.4%

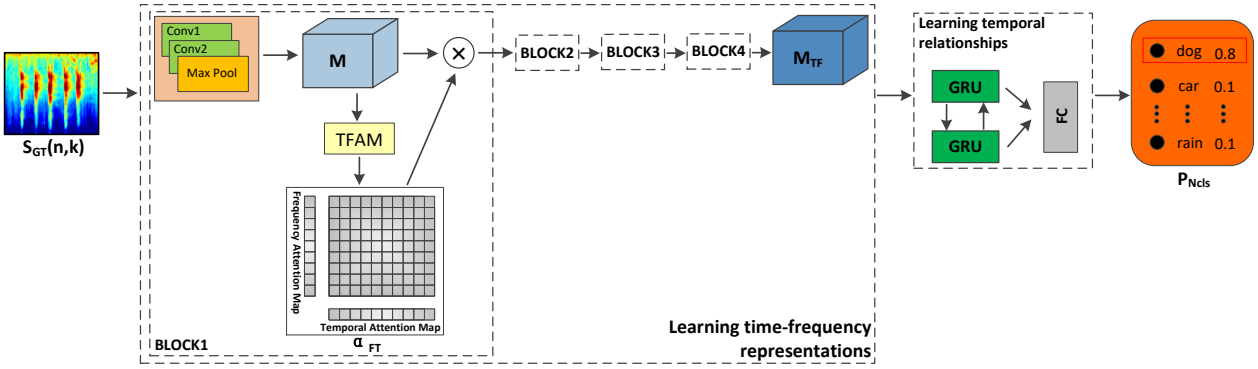


Figure 4. Illustration of our proposed environmental sound classification framework with temporal-frequency attention mechanism (TFAM).

6. Experiments

In order to prove the effectiveness of our proposed schemes, we numerically perform the experiments on a public environmental sound dataset called "ESC-50" [33] in this section. The ESC-50 dataset collects 2000 environmental recordings, which belong to 50 classes of 5 major categories, including animals, natural soundscapes and water sounds, human non-speech sounds, interior or domestic sounds, and exterior and urban noises. All audio samples are 5 s with a 44.1 kHz sampling frequency. In addition, all the experiments in this paper are obtained through five-fold cross-validation.

6.1. Experiment Setup

All the experiments are evaluated on the Nvidia P100 GPU for a fair comparison, and all models are trained by using Keras library with TensorFlow backend. In the training stage, we use the mini-batch stochastic gradient descent with Nesterov momentum of 0.9 and the learning rate scheme of reducing by 10 times per 100 epochs with initial value of 0.1. Moreover, we choose cross entropy as the loss function and the batch size is set as 200. Listed in Table 8 are some other important parameters.

Table 8. Parameter settings in experiments.

Parameters	Definition	Values
f_s	sampling frequency	44,100
N_{cls}	number of classes	50
T	STFT point	1024
N	frame length	128
K	number of Gammatone-filter banks	128
α	Mixup hyper-parameter	0.2

In the following scenarios, a simple CNN architecture shown in Figure 2 is use as a baseline, which models the relation between the log Gammatone spectrogram and the final results. Meanwhile, we numerically compare the classification performance of our proposed schemes with that of the baseline scheme.

6.2. Effect of Sub-Spectrogram Segmentation

We analyze the results under the influence of different N_{ss}, $\{f_i\}$, and $\{\omega^i\}$, as shown in Table 9. Firstly, we choose some f_i, and then we can attain a variety of different situations, such as different N_{ss} and the same N_{ss} with different $\{f_i\}$, by combining some of them. Then, we assign different $\{\omega^i\}$ to them and test the classification performance of the models.

The effect of N_{ss} in this system has been analyzed in Section 4.1, and the optimal number is $N_{ss}^{\star} = 4$. Here, we also compare the classification accuracy in three selection ways of $\{f_i\}$, including more segments in low-frequency portion, roughly average segmentation and more segments in high-frequency portion, and Table 9 shows the results, which can

prove that a higher classification accuracy can be obtained when more segments are applied in low-frequency portion.

Table 9. Classification accuracy under different N_{ss}, $\{f_i\}$, and $\{\omega^i\}$.

N_{ss}	f_L (kHz)	$\{f_i\}$ (kHz)	f_H (kHz)	$\{w^i\}$	Accuracy
1	0	-	22.05	1	77.9%
2	0	10	22.05	0.7, 0.3	79.9%
3	0	10, 20	22.05	0.5, 0.3, 0.2	80.2%
3	0	7, 14	22.05	0.5, 0.2, 0.3	80.6%
3	0	6, 10	22.05	0.5, 0.3, 0.2	81.7%
4	0	10, 15, 20	22.05	0.5, 0.2, 0.2, 0.1	80.3%
4	0	5, 10, 15	22.05	0.4, 0.3, 0.1, 0.2	80.9%
4	0	3, 6, 10	22.05	0.4, 0.2, 0.2, 0.2	82.1%
5	0	10, 13, 16, 19	22.05	0.4, 0.2, 0.1, 0.2, 0.1	81.0%
5	0	5, 10, 15, 20	22.05	0.4, 0.2, 0.1, 0.2, 0.1	80.7%
5	0	3, 6, 10, 15	22.05	0.4, 0.2, 0.2, 0.1, 0.1	81.8%
6	0	3, 6, 10, 13, 16	22.05	0.3, 0.2, 0.2, 0.1, 0.1, 0.1	81.3%
6	0	6, 10, 13, 16, 19	22.05	0.4, 0.1, 0.2, 0.1, 0.1, 0.1	81.2%

We can reach two conclusions by analyzing the curve in Figure 3, the one is that low-frequency band contains a large proportion of the characteristics of environmental sounds, and the other is that high-frequency band is still indispensable for ESC although it contains few of the characteristics. Therefore, in order to obtain better performance, we appropriately increased $\{\omega^i\}$ of low-frequency segments during fusion, and in Table 9, all $\{\omega^i\}$ are optimal in its corresponding situation.

6.3. Accuracy under Sub-Spectrogram Segmentation Based Classification Framework

Further, we compared the classification accuracy in different combinations of mixup, CNN, RNN, segmentation, and score level fusion. As shown in Table 10, the classification accuracy can be improved when using them together. Specifically, our highest classification accuracy is 82.1%, which has an absolutely improvement of 9.2% over the baseline system.

Table 10. Comparison for different combinations of mixup, CNN, RNN, segmentation, and score level fusion. When using segmentation, the N_{ss}, $\{f_i\}$ and $\{\omega^i\}$ are set as 4, {3, 6, 10} and {0.4, 0.2, 0.2, 0.2}, respectively.

Network	Mixup	Segmentation	Fusion	Accuracy
CNN	×	×	×	72.9%
CRNN	×	×	×	75.2%
CNN	√	×	×	76.1%
CRNN	√	×	×	77.9%
CNN	×	√	×	76.1%
CNN	×	√	√	77.6%
CRNN	×	√	×	78.1%
CRNN	×	√	√	79.5%
CNN	√	√	×	78.2%
CNN	√	√	√	80.8%
CRNN	√	√	×	79.8%
CRNN	√	√	√	82.1%

6.4. Accuracy under Temporal-Frequency Attention Based Classification Framework

We finally combine different strategies together to improve the overall classification accuracy, including CRNN architecture, different data augmentation schemes, as well as the proposed TFAM blocks. As shown in Table 11, by jointly utilizing all the above strategies, we can achieve a classification accuracy up to 86.4%, which corresponds to 3.9% improvement than sub-spectrogram segmentation based classification framework. This can also significantly demonstrate that the method we proposed is effective.

Table 11. Comparison for different combinations of CNN, RNN, miuxp, SpecAugment, segmentation, and TFAM.

Network	Mixup	SpecAugment	Segmentation	TFAM	Accuracy
CNN	×	×	×	×	72.9%
CRNN	×	×	×	×	75.2%
CRNN	√	×	×	×	77.9%
CRNN	×	√	×	×	77.5%
CRNN	×	×	√	×	79.5%
CRNN	×	×	×	√	80.3%
CRNN	√	×	√	×	82.1%
CRNN	√	×	×	√	83.1%
CRNN	×	√	×	√	82.7%
CRNN	√	√	×	√	86.4%

In addition, Figure 5 shows the confusion matrix when the classification accuracy is 86.4%. The confusion matrix fully displays the correctness and wrongness of the classification of each class. The horizontal axis and vertical axis in the figure, respectively, represent the predicted labels and the true labels of 50 classes of environmental sound. Among them, the corresponding relationship between the number from 1 to 50 and the environmental sound class is: 1—Dog; 2—Rooster; 3—Pig; 4—Cow; 5—Frog; 6—Cat; 7—Hen; 8—Insect (flying); 9—Sheep; 10—Crow; 11—Rain; 12—Sea waves; 13—Crackling fire; 14—Crickets; 15—Chirping birds; 16—Water drops; 17—Wind; 18—Pouring water; 19—Toilet flush; 20—Thunderstorm; 21—Crying baby; 22—Sneezing; 23—Clapping; 24—Breathing; 25—Coughing; 26—Footsteps; 27—Laughing; 28—Brushing teeth; 29—Snoring; 30—Drinking, sipping; 31—Door knock; 32—Mouse click; 33—Keyboard typing; 34—Door, wood creaks; 35—Can opening; 36—Washing machine; 37—Vacuum cleaner; 38—Clock alarm; 39—Clock tick; 40—Glass breaking; 41—Helicopter; 42—Chainsaw; 43—Siren; 44—Car horn; 45—Engine; 46—Train; 47—Church bells; 48—Airplane; 49—Fireworks; 50—Hand saw.

Finally, we compare the classification accuracy of the method proposed in this paper with the existing methods, as shown in Table 12. It can be seen that compared with most existing methods, the proposed method has obvious advantages in classification accuracy.

Table 12. Classification accuracy comparison.

Existing Methods	TFAM
PiczakCNN	64.9%
Google Net	67.8%
SoundNet	74.2%
AlexNet	78.7%
WaveMsNet	79.1%
ProCNN	82.8%
Multi-Stream CNN	83.5%
EnvNet-v2	84.9%
ACRNN	86.1%
TFAM	86.4%

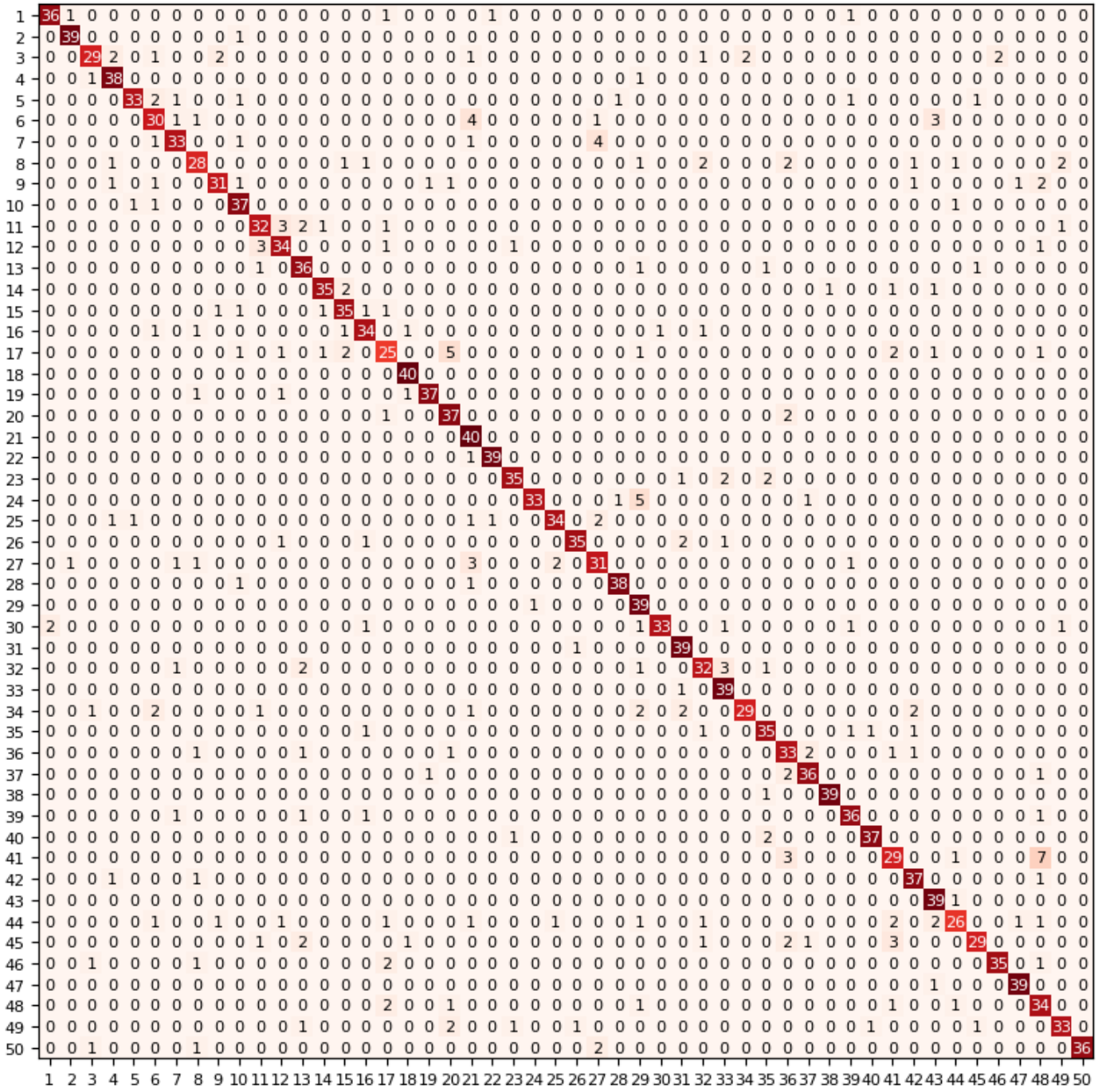


Figure 5. The confusion matrix when the classification accuracy is 86.4%.

7. Conclusions

In this paper, we have successively proposed two effective environmental sound classification frameworks based on sub-spectrogram segmentation and temporal-frequency domain attention. The proposed frameworks jointly consider the recurrent network architecture, the data augmentation policies, as well as feature enhancement schemes to improve the classification accuracy of ESC-50. Numerical results show that our proposed frameworks can achieve 82.1% and 86.4% classification accuracy on ESC-50 dataset, respectively, which is equivalent to more than 13.5% improvement over the traditional baseline scheme.

Author Contributions: Conceptualization, T.Q.; Data curation, T.Q. and S.Z.; Funding acquisition, S.C. and S.X.; Investigation, T.Q.; Methodology, T.Q.; Project administration, S.Z.; Resources, S.C. and S.X.; Supervision, S.Z. and S.X.; Writing—original draft, T.Q.; Writing—review & editing, S.Z., S.C. and S.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Key-Area Research and Development Program of Guangdong Province(2020B0101130012), Foshan Science and Technology Innovation Team Project (FS0AA-KJ919-4402-0060), the National Natural Science Foundation of China (NSFC) under Grants 62071284, 61871262, 61901251 and 61904101, the National Key Research and Development Program of China under Grants 2017YEF0121400 and 2019YFE0196600, the Innovation Program of Shanghai Municipal Science and Technology Commission under Grant 20JC1416400, and research funds from Shanghai Institute for Advanced Communication and Data Science (SICS).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This work was supported by Key-Area Research and Development Program of Guangdong Province(2020B0101130012), Foshan Science and Technology Innovation Team Project (FS0AA-KJ919-4402-0060), the National Natural Science Foundation of China (NSFC) under Grants 62071284, 61871262, 61901251 and 61904101, the National Key Research and Development Program of China under Grants 2017YEF0121400 and 2019YFE0196600, the Innovation Program of Shanghai Municipal Science and Technology Commission under Grant 20JC1416400, and research funds from Shanghai Institute for Advanced Communication and Data Science (SICS).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barchiesi, D.; Giannoulis, D.; Stowell, D.; Plumbley, M.D. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Process. Mag.* **2015**, *32*, 16–34. [CrossRef]
2. Zatorre, R.J.; Belin, P.; Penhune, V.B. Structure and function of auditory cortex: Music and speech. *Trends Cogn. Sci.* **2002**, *6*, 37–46. [CrossRef]
3. Chachada, S.; Kuo, C.C.J. Environmental sound recognition: A survey. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, e14. [CrossRef]
4. Chu, S.; Narayanan, S.; Kuo, C.C.J. Environmental sound recognition with time–frequency audio features. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1142–1158. [CrossRef]
5. Cowling, M.; Sitte, R. Comparison of techniques for environmental sound recognition. *Pattern Recogn. Lett.* **2003**, *24*, 2895–2907. [CrossRef]
6. Jalil, M.; Butt, F.A.; Malik, A. Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In Proceedings of the International Conference on TAEECE, Konya, Turkey, 9–11 May 2013; pp. 208–212.
7. Dennis, J.W. *Sound Event Recognition in Unstructured Environments Using Spectrogram Image Processing*; Nanyang Technological University: Singapore, 2014.
8. Rabiner, L.R.; Juang, B.H.; Rutledge, J.C. *Fundamentals of Speech Recognition*; Prentice-Hall: Hoboken, NJ, USA, 1993; Volume 14.
9. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 25th International Workshop Machine Learning Signal Processing, Boston, MA, USA, 17–20 September 2015; pp. 1–6.
10. Valero, X.; Alias, F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Trans. MultiMedia* **2012**, *14*, 1684–1689. [CrossRef]
11. Keller, J.M.; Gray, M.R.; Givens, J.A. A fuzzy k-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybern.* **1985**, *4*, 580–585. [CrossRef]
12. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
13. Scholkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2001.
14. Atrey, P.K.; Maddage, N.C.; Kankanhalli, M.S. Audio based event detection for multimedia surveillance. In Proceedings of the 2006 International Conference on Acoustics Speech and Signal Process, Toulouse, France, 14–19 May 2006; Volume 5, p. V.
15. Bisot, V.; Serizel, R.; Essid, S.; Richard, G. Feature learning with matrix factorization applied to acoustic scene classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1216–1229. [CrossRef]
16. Tokozume, Y.; Ushiku, Y.; Harada, T. Learning from between-class examples for deep sound recognition. *arXiv* **2018**, arXiv:1711.10282.
17. McLoughlin, I.; Zhang, H.; Xie, Z.; Song, Y.; Xiao, W. Robust sound event classification using deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 540–552. [CrossRef]

18. Boddapati, V.; Petef, A.; Rasmusson, J.; Lundberg, L. Classifying environmental sounds using image recognition networks. *Procedia Comput. Sci.* **2017**, *112*, 2048–2056. [CrossRef]
19. Zhang, X.; Zou, Y.; Shi, W. Dilated convolution neural network with LeakyReLU for environmental sound classification. In Proceedings of the 22nd International Conference on Digital Signal Processing, London, UK, 23–25 August 2017; pp. 1–5.
20. Phaye, S.S.R.; Benetos, E.; Wang, Y. SubSpectralNet-Using Sub-Spectrogram based Convolutional Neural Networks for Acoustic Scene Classification. *arXiv* **2018**, arXiv:1810.12642.
21. Guo, J.; Xu, N.; Li, L.J.; Alwan, A. Attention Based CLDNNs for Short-Duration Acoustic Scene Classification. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 469–473.
22. Jun, W.; Shengchen, L. Self-Attention Mechanism Based System for Dcase2018 Challenge Task1 and Task4. In Proceedings of the DCASE Challenge, Surrey, UK, 19–20 November 2018.
23. Zhang, Z.; Xu, S.; Qiao, T.; Zhang, S.; Cao, S. Attention based Convolutional Recurrent Neural Network for Environmental Sound Classification. *arXiv* **2019**, arXiv:1907.02230.
24. Li, X.; Chebiyyam, V.; Kirchhoff, K. Multi-stream Network with Temporal Attention For Environmental Sound Classification. *arXiv* **2019**, arXiv:1901.08608.
25. Ren, Z.; Kong, Q.; Qian, K.; Plumbley, M.D.; Schuller, B. Attention-Based Convolutional Neural Networks for Acoustic Scene Classification. In Proceedings of the DCASE Challenge, Surrey, UK, 19–20 November 2018.
26. Wang, H.; Zou, Y.; Chong, D.; Wang, W. Learning discriminative and robust time-frequency representations for environmental sound classification. *arXiv* **2019**, arXiv:1912.06808.
27. Zhang, Z.; Xu, S.; Zhang, S.; Qiao, T.; Cao, S. Learning Attentive Representations for Environmental Sound Classification. *IEEE Access* **2019**, *7*, 130327–130339. [CrossRef]
28. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
29. Zhang, Z.; Xu, S.; Cao, S.; Zhang, S. Deep convolutional neural network with mixup for environmental sound classification. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Guangzhou, China, 23–26 November 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 356–367.
30. Suh, S.; Lim, W.; Park, S.; Jeong, Y. Acoustic Scene Classification Using SpecAugment and Convolutional Neural Network with Inception Modules. In Proceedings of the DCASE2019 Challenge, New York, NY, USA, 25–26 October 2019.
31. Qiao, T.; Zhang, S.; Zhang, Z.; Cao, S.; Xu, S. Sub-Spectrogram Segmentation for Environmental Sound Classification via Convolutional Recurrent Neural Network and Score Level Fusion. In Proceedings of the 2019 IEEE International Workshop on Signal Processing Systems (SiPS), Nanjing, China, 20–23 October 2019; pp. 318–323.
32. Van Loan, C.F. The ubiquitous Kronecker product. *J. Comput. Appl. Math.* **2000**, *123*, 85–100. [CrossRef]
33. Piczak, K.J. ESC: Dataset for environmental sound classification. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1015–1018.
34. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
35. Wang, H.; Zou, Y.; Chong, D.; Wang, W. Environmental Sound Classification with Parallel Temporal-spectral Attention. *arXiv* **2020**, arXiv:1912.06808v3.
36. Phan, H.; Chén, O.Y.; Pham, L.; Koch, P.; De Vos, M.; McLoughlin, I.; Mertins, A. Spatio-temporal attention pooling for audio scene classification. *arXiv* **2019**, arXiv:1904.03543.
37. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.