2018 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP2018)
Honolulu, Hawaii, USA, March 4-7, 2018

# Acoustic Scene Classification Based on Spatial Feature Extraction Using Convolutional Neural Networks

Gen Takahashi, Takeshi Yamada, and Shoji Makino

University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
E-mail: g.takahashi@mmlab.cs.tsukuba.ac.jp

## Abstract

Acoustic scene classification (ASC) classifies the place or situation where an acoustic sound was recorded. The Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge prepared a task involving ASC. Some methods using convolutional neural networks (CNNs) were proposed in the DCASE 2017 Challenge. The best method independently performed convolution operations for the left, right, mid (addition of left and right channels), and side (subtraction of left and right channels) input channels to capture spatial features. On the other hand, we propose a new method of spatial feature extraction using CNNs. In the proposed method, convolutions are performed for the time-space (channel) domain and frequency-space domain in addition to the time-frequency domain to capture spatial features. We evaluate the effectiveness of the proposed method using the task in the DCASE 2017 Challenge. The experimental results confirmed that convolution operations for the frequency-space domain are effective for capturing spatial features. Furthermore, by using a combination of the three domains, the classification accuracy was improved by 2.19% compared with that obtained using the time-frequency domain only.

## 1. Introduction

Attempts have been made to automatically recognize human behavior and surrounding circumstances. This technology is applicable to the monitoring of elderly people, the auto-tagging of multimedia contents, and life log collection. Acoustic event detection (AED) and acoustic scene classification (ASC) have been focused on as fundamental technologies used in these systems. AED detects acoustic events including sound signals and their timestamps, where an acoustic event is a single sound emitted from one sound source, such as door opening, coughing, or toppling. On the other hand, ASC classifies the place or situation where the acoustic sound was recorded. The length of an acoustic sound is typically on the order of 10 s, and types of sounds include buses, office noise, and home noise. In this paper, we focus on ASC.

Recently, ASC has been actively investigated. For example, the Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge [1] was held and a task involving ASC was prepared for it. Some methods using convolutional neural networks (CNNs) were proposed in the DCASE 2017 Challenge [2][3]. The method in [3] independently performed convolutions for the left, right, mid (addition of left and right channels), and side (subtraction of left and right channels) input channels to capture spatial features.

On the other hand, we propose a new method of spatial feature extraction using CNNs. In the proposed method, convolutions are performed for the time-space (channel) domain and frequency-space domain in addition to the time-frequency domain to capture spatial features. We evaluate the effectiveness of the proposed method using the task in the DCASE 2017 Challenge.

## 2. Proposed method

### 2.1 Process flow of the proposed method

Figure 1 shows the process flow of the proposed method. We assume two input channels because the acoustic data in the DCASE 2017 Challenge had two channels, and that the length of an acoustic sound is 10 s.

First, we compute 128th-order log mel filter bank outputs for both the left and right channels. The time frame length and frame shift length in the frame analysis are 40 and 20 ms, respectively. Then, these time series of features are divided into blocks of ten frames to be input to CNNs. The reason for dividing features into blocks is that it was reported that features in temporally too distant frames do not contribute to improving classification accuracy [4]. Finally, we perform ASC by computing the logarithmic sum of the output probabilities of the CNNs for each block for each acoustic scene. We describe the spatial feature extraction using CNNs below.

### 2.2 Spatial feature extraction using CNNs

In ASC, it is important to use spatial information. For example, simply changing from a mono channel to stereo channels leads to an improvement of classification accuracy. However, in ASC, conventional CNNs are applied to the time-frequency domain. Therefore, in the proposed method, we extract features containing spatial information by performing convolutions for the time-space (stereo) domain and frequency-space domain in addition to the time-frequency domain.

The top, center, and bottom parts of Figure 1 represent the CNNs being applied to the time-frequency domain, time-

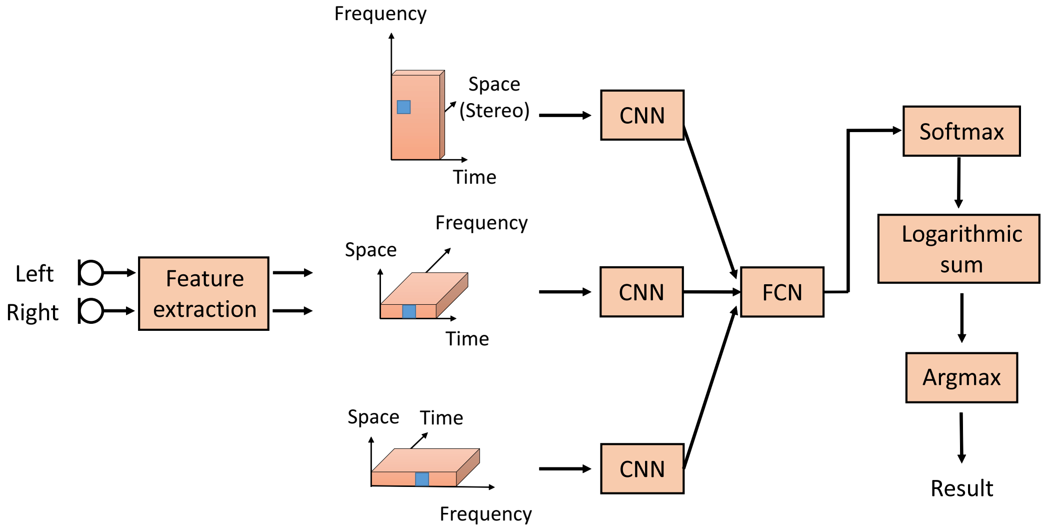
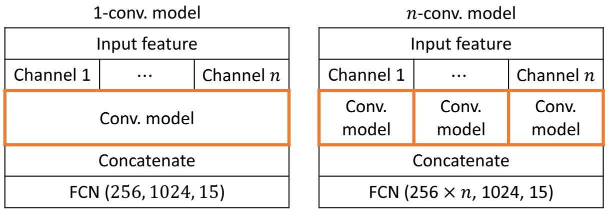


Figure 1: Process flow of the proposed method



| 1-conv. model | | |
|---|---|---|
| Input feature | | |
| Channel 1 | ··· | Channel $n$ |
| Conv. model | | |
| Concatenate | | |
| FCN (256, 1024, 15) | | |

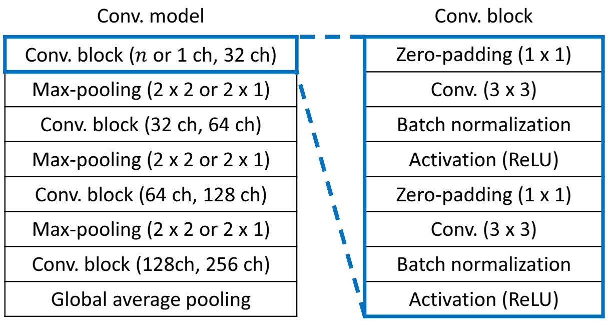
| $n$-conv. model | | |
|---|---|---|
| Input feature | | |
| Channel 1 | ··· | Channel $n$ |
| Conv. model | Conv. model | Conv. model |
| Concatenate | | |
| FCN (256 × $n$, 1024, 15) | | |

Figure 2: Overview of the CNNs in the proposed method

space domain, and frequency-space domain, respectively. Also, blue squares represent filters when convolutions are performed. For example, in the convolution for the time-frequency domain, the time-frequency spectrogram is regarded as a feature map and space (stereo) is regarded as the channels of the CNNs. Therefore, the number of channels of the CNNs is two. On the other hand, in the convolution for the time-space domain, since frequency is regarded as the channel of the CNNs, the number of channels of the CNNs is 128. Convolutions for the time-space domain or frequency-space domain correspond to extracting spatial information between stereo channels in a feature map.

### 2.3 Network architecture

In the proposed method, we used CNNs consisting of eight convolution layers on the basis of [3]. Figure 2 shows an overview of the CNNs in the proposed method. First, we explain the 1-convolution model on the left side of the figure. In the 1-conv. model, one conv. model is used. The input of the conv. model is a feature map of $n$ channels, and the output is a feature map of $1 \times 1 \times 256$. The conv. model is described later. Then, the output of the conv. model is concatenated to obtained 256th-order features. Finally, fully connected neural networks (FCNs) consisting of 1024 and 15 units are applied to the features and the output probability is obtained. Next, we explain the $n$-conv. model. In the $n$-conv. model, $n$ conv. models are used. The input of the conv. model is a feature map of 1 channel, and the output is a feature map of $1 \times 1 \times 256$. Also, one conv. model is applied independently to each channel. Therefore, the dimension of the feature obtained by concatenating the output of each conv. model is $256 \times n$. The subsequent process is the same as that for the 1-conv. model. Since each channel is inputted independently when applying the CNNs in the $n$-conv. model, an $n$-conv. model can extract the characteristics of each channel. On the other hand, the number of parameters in CNNs becomes about $n$ times as large as that for the 1-conv. model.



| Conv. model |
|---|
| Conv. block ($n$ or 1 ch, 32 ch) |
| Max-pooling (2 x 2 or 2 x 1) |
| Conv. block (32 ch, 64 ch) |
| Max-pooling (2 x 2 or 2 x 1) |
| Conv. block (64 ch, 128 ch) |
| Max-pooling (2 x 2 or 2 x 1) |
| Conv. block (128ch, 256 ch) |
| Global average pooling |

| Conv. block |
|---|
| Zero-padding (1 x 1) |
| Conv. (3 x 3) |
| Batch normalization |
| Activation (ReLU) |
| Zero-padding (1 x 1) |
| Conv. (3 x 3) |
| Batch normalization |
| Activation (ReLU) |

Figure 3: Conv. model and conv. block

Next, we explain the conv. model. Figure 3 shows the conv. model and the conv. block used in the conv. model. In the conv. model, the conv. block that performs convolutions is repeatedly alternated with max-pooling. The two numbers

Table 1: Overview of the development dataset

| | |
|---|---|
| # of scenes | 15 |
| # of sound data | 4680<br>(=15 scenes × 312 data) |
| data length | 10 s |
| # of channels | 2 (left and right) |
| sampling frequency | 44.1 kHz |
| quantization bits | 24 bit |

Table 2: Conditions of the acoustic features and the CNNs

| | |
|---|---|
| feature | 128th-order<br>log mel filter bank outputs |
| frame length | 40 ms |
| frame shift length | 20 ms |
| audio block size | 10 frames (without overlap) |
| domain | time-frequency,<br>time-space,<br>frequency-space and<br>combination of 3 domains |
| # of conv. layers | 8 |
| optimization method | Adam |
| epoch | 20 |

Table 3: Classification accuracies of proposed methods

| domain \ model | 1-conv. | $n$-conv. |
|---|---|---|
| time-frequency domain | 81.95 | 80.01 |
| time-space domain | 78.81 | 80.84 |
| frequency-space domain | 82.20 | 83.55 |
| combination of 3 domains | 81.69 | 84.14 |

separated by a comma in the parentheses of a conv. block are the number of input channels of the convolution layer on the left and the number of output channels on the right. The number of input channels of the first conv. block is $n$ for the 1-conv. model and 1 for the $n$-conv. model. The numbers in the parentheses of the pooling layer are the window size used for pooling. The window size in the pooling layer is $2 \times 2$ for the time-frequency domain. On the other hand, for the time-space domain or frequency-space domain, since the dimension of the space is only two, the window size in the pooling layer is $2 \times 1$. In the final pooling layer, average pooling is performed for the entire feature map. Therefore, the dimension of the feature map after average pooling is $1 \times 1 \times 256$ (the number of channels).

Finally, we explain the conv. block. In the conv. block, zero padding with a size of $1 \times 1$ is first performed for the input features. Next, a convolution with filters of size $3 \times 3$ is performed. Then, batch normalization [5] is performed. Finally, the ReLU function is applied as an activation function. These processes are performed twice in the conv. block. In this paper, we classified sound scenes using the 1-conv. model and n-conv. model for each domain and compared the results.

## 3. Evaluation

### 3.1 Experimental conditions

In this section, we evaluate the effectiveness of spatial feature extraction by the proposed method using the development dataset of the DCASE 2017 Challenge. Table 1 shows an overview of the development dataset. The dataset contains 15 acoustic scenes (see Table 4). Each scene has 312 sound data, each of which is a stereo signal with a duration of 10 s. The sampling frequency is 44.1 kHz and the number of quantization bits is 24.

Table 2 shows the conditions of the acoustic features and the CNNs. The features are 128th-order log mel filter bank outputs. The time frame length and frame shift length in the frame analysis are 40 and 20 ms, respectively, which are the same as those used in the baseline system of the DCASE 2017 Challenge. The block size when dividing the time series of features is 10 frames. Adam [6] is used as the optimization method for training a classifier, and the number of epochs during training is set to 20.

The classifier is evaluated using the average of the classification accuracies obtained by fourfold cross-validation (data open, acoustic scene closed). The way of dividing the data in the fourfold cross-validation is the same as that in the DCASE 2017 Challenge. We used Chainer [7] to implement the proposed method.

In this experiment, we apply two models, the 1-conv. model and the $n$-conv. model, to the following four domains:

- time-frequency domain only,
- time-space domain only,
- frequency-space domain only, and
- a combination of the above three domains.

### 3.2 Results

Table 3 shows the classification accuracy for each pair of domains and models. The rows are the domains where convolutions are performed, and the columns are the models of CNNs. Comparing the classification accuracies of each domain, the classification accuracy in the frequency-space domain (third row) was the highest among the three domains. This means that better spatial features can be extracted by using the frequency-space domain than by using the time-frequency domain (first row). On the other hand, little improvement in the classification accuracy was obtained in the time-space domain (second row) including the space. Next, comparing the classification accuracies of each model, the classification accuracies in $n$-conv. model were improved in

Table 4: Classification accuracies of each sound scene for proposed methods

| domain | time-frequency | | time-space | | frequency-space | | combination | |
|---|---|---|---|---|---|---|---|---|
| model | 1-conv. | n-conv. | 1-conv. | n-conv. | 1-conv. | n-conv. | 1-conv. | n-conv. |
| bus | 89.74% | 79.49% | 76.92% | 91.99% | 86.86% | 91.35% | 84.94% | 83.97% |
| café/restaurant | 62.18% | 51.28% | 42.31% | 38.46% | 47.44% | 37.18% | 61.86% | 47.44% |
| car | 92.63% | 97.12% | 86.22% | 88.14% | 95.19% | 97.44% | 90.06% | 96.79% |
| city_center | 95.83% | 85.26% | 90.38% | 96.15% | 85.90% | 95.19% | 86.86% | 82.37% |
| forest_path | 95.19% | 93.27% | 78.85% | 75.96% | 96.79% | 92.63% | 92.95% | 93.91% |
| grocery_store | 70.83% | 60.58% | 77.24% | 88.14% | 77.56% | 91.67% | 70.51% | 84.62% |
| home | 75.79% | 80.82% | 78.62% | 77.36% | 75.16% | 83.33% | 77.04% | 77.99% |
| beach | 85.26% | 70.51% | 80.77% | 80.77% | 87.18% | 89.42% | 85.90% | 83.01% |
| library | 86.22% | 77.56% | 82.37% | 77.56% | 83.65% | 83.01% | 80.13% | 88.14% |
| metro_station | 95.83% | 95.51% | 97.12% | 98.72% | 98.08% | 96.79% | 93.59% | 97.12% |
| office | 91.03% | 99.68% | 98.08% | 98.40% | 91.99% | 97.12% | 99.04% | 98.72% |
| residential_area | 70.19% | 86.22% | 80.13% | 76.60% | 78.21% | 78.85% | 83.01% | 84.62% |
| train | 68.59% | 70.51% | 59.94% | 65.71% | 75.00% | 69.23% | 71.47% | 78.85% |
| tram | 82.37% | 89.10% | 85.58% | 90.06% | 86.86% | 88.78% | 92.63% | 88.78% |
| park | 67.63% | 63.14% | 67.63% | 68.59% | 67.31% | 61.22% | 55.45% | 75.96% |

the domain including the space compared with those in the 1-conv. model: however, it was not improved in the time-frequency domain. It is considered that when convolutions are performed independently for each channel for the time-frequency domain, spatial information is lost in the calculation process.

The highest classification accuracy was obtained by using the combination of three domains (fourth row) and the $n$-conv. model. It was improved by 2.19% compared with that for the time-frequency domain and 1-conv. model. It was also improved by 9.34% compared with that for the baseline system of the DCASE 2017 Challenge. For the baseline system, the input features are 40th-order mel filter bank outputs, and the classifier is an FCN with two hidden layers and 50 hidden units.

Finally, Table 4 shows the classification accuracy for each acoustic scene for each domain and model. Orange cells represent the highest classification accuracy among the eight methods for each scene. The number of orange cells in the frequency-space domain was larger than that in the time-frequency domain. It was found that the classification accuracy of "grocery store" and "home" was improved greatly compared with that for the time-space domain and 1-conv. model. Also, by using a combination of three domains, the classification accuracies of scenes with low classification accuracies in the time-space domain and 1-conv. model such as "grocery store", "home", "residential area", "train", and "park" were improved. On the other hand, the classification accuracy of "cafe/restaurant" degraded. Therefore, it is necessary to investigate the scenes for which the classification accuracy degraded and to improve the classification accuracies of those scenes in future.

## 4. Conclusions

In this paper, we proposed a spatial feature extraction method using CNNs. We carried out an experiment using the DCASE 2017 Challenge development dataset and evaluated the effectiveness of spatial feature extraction using the proposed method. It was found that by applying the $n$-conv. model to the frequency-space domain, the classification accuracy was improved compared with that obtained by applying the 1-conv. model to the time-frequency domain. Furthermore, by using a combination of three domains, the classification accuracy was improved by 2.19% compared with that obtained using the time-frequency domain only.

## References

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj and T. Virtanen, "DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System," DCASE 2017 Challenge.

[2] S. Mun, S. Park, D. Han and H. Ko, "Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane," DCASE 2017 Challenge.

[3] Y. Han and J. Park, "Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification," DCASE 2017 Challenge.

[4] G. Takahashi, T. Yamada, N. Ono and S. Makino, "Performance evaluation of acoustic scene classification using DNN-GMM and frame-concatenated acoustic features," Proc. APSIPA 2017, Paper ID 219, December 2017.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," ArXiv e-prints, February 2015.

[6] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in Proceedings of International Conference on Learning Representations, 2015.

[7] https://chainer.org/