PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

# A Study on the Data Augment Method considering Room Transfer Functions for Acoustic Scene Classification

Minhan Kim[1]; Seokjin Lee[2]

[1, 2] Kyoungpook National University, Korea

## ABSTRACT

Acoustic scene classification is the problem of recognition of sound around our living area. Since people recognize the situation through sound when they can't see, it is very natural that acoustical approach is being made in research that awareness of environment. In this field, research using deep learning method such as CNN is widely used recently. However, this method has the disadvantage that the lower the number of data, the lower the performance. So, in this paper, data augmentation considering acoustical approach-the transfer function of the room-was performed to obtain enough number of data for each classes. To verify this method, we used dataset from DCASE 2018 challenge, which is acoustic scene classification competition. Our augmentation method improved overall f1-score by 0.1 from the state-of-art performance.

Keywords: Acoustic Scene Classification, Data Augmentation, Machine Learning

## 1. INTRODUCTION

Acoustic Scene Classification (ASC) is field of recognize sound events in user environment sounds. It has been actively studied to be used as an acoustic sensors like sensors of other domains in order to applied to areas such as context aware computing(1).

Currently, a machine learning based approach is mainly used to deal with this field. Since machine learning requires a large amount of data to develop optimized model, numerous training data which is related with target data is needed. However, collecting a sufficient number of labeled data is difficult because collecting data is a costly and time-consuming process. Therefore, data augmentation technique which use computing power to make artificial data is being used in process of training model of machine learning.

The most commonly used techniques for data augmentation are adding noise(2), time shifting(3) and mix and shuffle(4). Although this traditional methods contributes to improving the performance of the model, it is more intuitive whether than the acoustical approach. So far, data augmentation considering acoustical approach have been studied, but they mainly focused on speech recognition problem(5).

In this study, data augmentation considering acoustical approach-the transfer function of the room-was performed to obtain enough number of data for training model in ASC. The sound which is captured from acoustic sensors has transformed through the room transfer function. For that reason, we produced data that has different room transfer function by multiplying Mel-spectrogram of sound source to random number that has Gaussian distribution.

## 2. PROBLEM DESCRIPTION

### 2.1 Acoustic scene classification

The goal of this field is to collect sound from acoustic devices such as microphone and classify sound event through the classifier, as shown in Figure 1.

[1] kmh7576@naver.com
[2] sjlee6@knu.ac.kr


**Figure 1 Conceptual overview of the ASC**

## 2.2 Approaches

Currently, approaches to solving this problem are mainly using machine learning. The train and evaluation dataset train the model and uses weights determined to predict unknown signals. Figure 2 shows the overall process of machine learning.

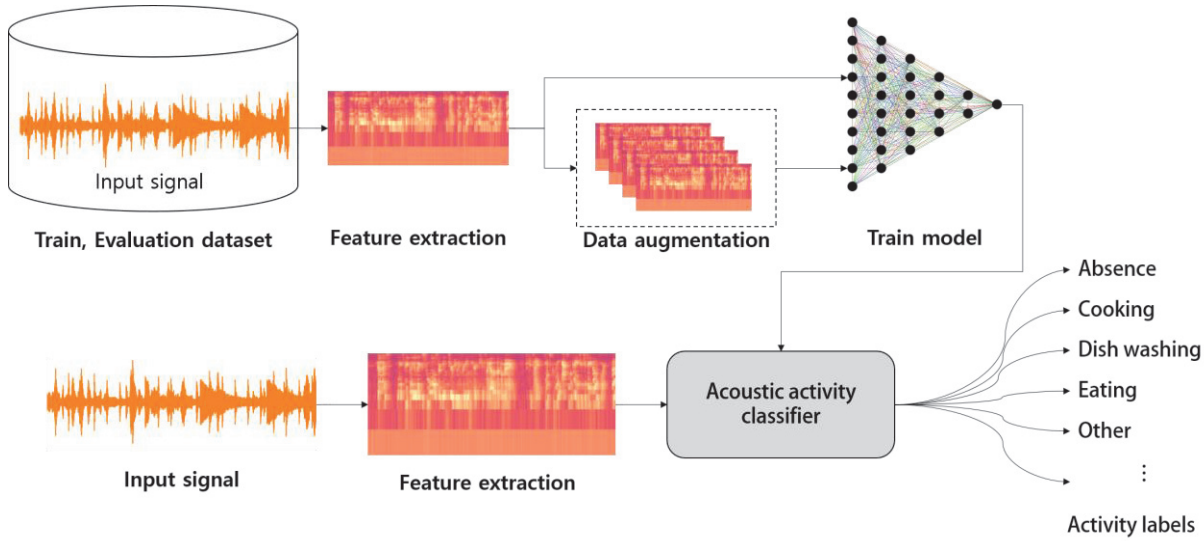


**Figure 2 Machine learning process**

In this process, the part that determines the performance of the model is roughly the relevance between the data to be trained and the data to be predicted, the feature selection, and the structure of the model. We will focus on data relevance.

As mentioned in introduction, data augmentation methods using given data are mainly used. In terms of data relevance, the data which has similarity with data to predict is required. For that purpose, we used data augmentation technique to increase the number of data in a fewer labels.

## 3. DATA AUGMENTAION

As shown in Figure 3, we hear the sound from the sound source through the room transfer function.

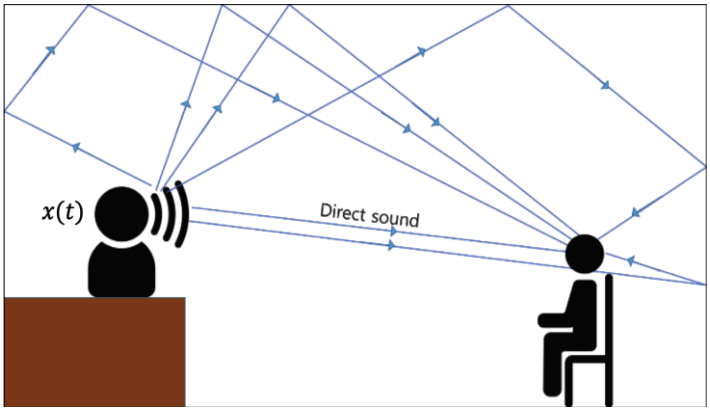


**Figure 3 Sound transmission path**

The above situation can be expressed by

$$y_1(t) = h_1(t) * x(t) \tag{1}$$

where $x(t)$ is sound source, $h_1(t)$ is room transfer function and $y_1(t)$ is sound signal captured by acoustic sensors. By applying Fourier transform, eq. (1) can be easily represented in frequency domain:

$$Y_1(f) = H_1(f)X(f) \tag{2}$$

We can assume that same sound source x(t) has room transfer function $h_2(t)$ which is different from $y_1(t)$ then signal captured by acoustic sensor will be $y_2(t)$. With Fourier transformation we can express:

$$Y_2(f) = H_2(f)X(f) \tag{3}$$

Therefore by dividing eq. (2) with eq. (3) we can get:

$$Y_2(f) = Y_1(f)\left[\frac{H_2(f)}{H_1(f)}\right] \tag{4}$$

In the frequency domain as shown in eq. (4), the unknown signal $Y_2(f)$ can be obtained by dividing $H_2(f)$ by $H_1(f)$. Here we assumed that this value ($\left[\frac{H_2(f)}{H_1(f)}\right]$) is a random number with a Gaussian distribution.

# 4. EXPERIMENTS

## 4.1 Settings

Data augmentation was tested in DCASE 2018 Challenge Task5(6). Dataset is from SINS dataset(7) which is for detection of daily activities in a home environment using an acoustic sensor network. Figure 2 shows floorplan of the SINS dataset environment with sensor position.

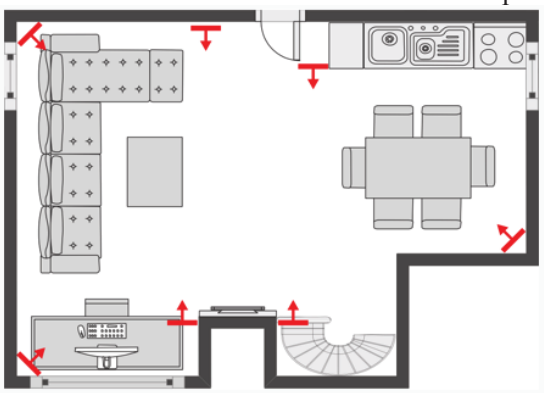

**Figure 4 2D floorplan of the combined kitchen and living room with the used sensor nodes.**

The continuous recordings from sensor nodes were split into audio segment of 10 seconds. Each sensor node has 4 microphones.

To predict signal, we built model by using machine learning as described in section 2.2. The experiments are carried out using the 4-fold cross validation setting in development dataset of DCASE 2018 Challenge. We applied data augmentation to each fold and utilized ADAM optimizer(8)with initial learning rate 0.0001 and a batch size of 128 samples. On training epoch, we chose network weight which is in the best accuracy on validation data.

## 4.2 Feature extraction

The split audio segment was feature extracted using Mel-spectrogram. Specification of Mel feature is shown in table 1.

**Table 1 Mel Spectrogram Specification**

| | |
|---|---|
| Number of Mel | 40 |
| Number of FFT | 1024 |
| Sample rate | 16000 |

## 4.3 Data augmentation

Data augmentation was applied to train dataset. Figure3 shows number of data by label and the number of augmented data.


**Figure 5 Number of data and augmented data in train dataset**

As can be seen in the figure 5, the data augmentation was applied to six labels with a small number of data.

### 4.4 Model architecture

To build our model, we applied same network architecture as the model of Inoue et al.(4). The network architecture and parameters are shown in Table 2.

**Table 2 Model architecture**

| Layer | Output size |
|---|---|
| Input | 40 × 501 × 1 |
| Conv(7 × 1, 64) + BN + ReLU | 40 × 501 × 64 |
| Max pooling(4 × 1) + Dropout(0.2) | 10 × 501 × 64 |
| Conv(10 × 1, 128) + BN + ReLU | 1 × 501 × 128 |
| Conv(1 × 7, 256) + BN + ReLU | 1 × 501 × 256 |
| Global max pooling + Dropout(0.5) | 256 |
| Dense | 128 |
| Softmax output | 9 |

### 4.5 EXPERIMENTAL RESULTS

Figure 4 shows F1 scores on each label and overall compared with traditional data augmentation method.

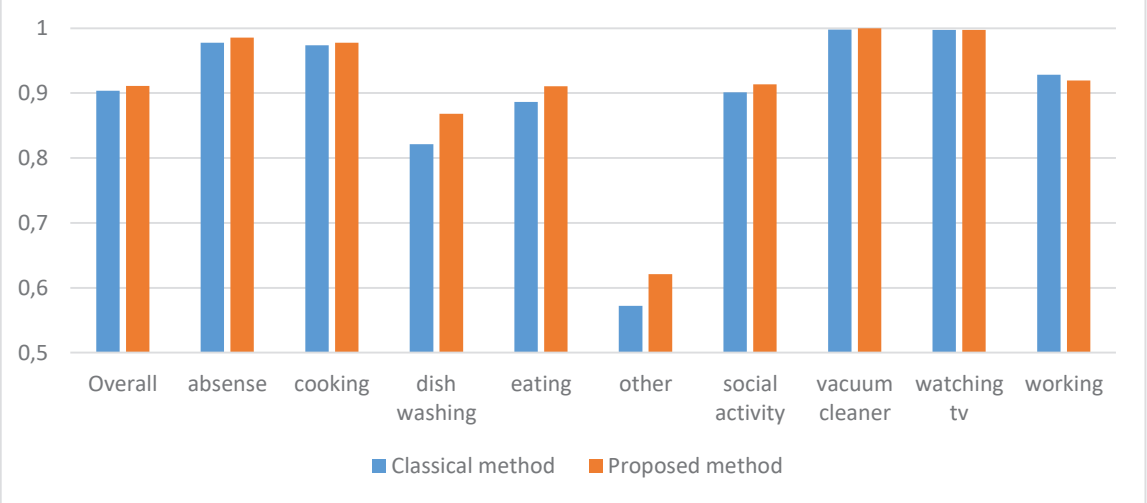


**Figure 6 F1 scores on each label**

As shown in figure 6, data augmentation using room transfer function enhanced accuracy of the model. The overall F1 score in proposed method is 0.91 while classical method is 0.9. In particular, we can see that the proposed method effectively improves the F1 score of the label with a relatively large number of data like 'dish washing', 'eating', 'other' and 'social activity'. We can see that the f1-score of the working label is slightly lowered, which can be thought of as performance deviation due to model learning.

## 5. CONCLUSIONS

In this paper, we considered improving ASC performance with using machine learning technique. Specifically, we took advantage of the fact that machine learning requires large amounts of data to achieve a certain level of performance. From this point of view, we proposed data augmentation method with considering room transfer function. To confirm the performance, we utilized DCASE 2018 Challenge Task5. The result of experiment shows that our method enhanced machine learning performance particularly in labels which are lower number of data than others.

## ACKNOWLEDGEMENTS

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding)

## REFERENCES

[1] Chu, S; Narayanan, S; Kuo, CJ. Time – frequency audio features. Ieee transactions on audio, speech, and language processing. 17(6), 2009, 1142–58.

[2] Yin, S; Liu, C; Zhang, Z; Lin, Y; Wang, D; Tejedor, J; et al. Noisy training for deep neural networks in speech recognition. Eurasip J Audio, Speech, Music Process. 2015(1), 2015, 1–14.

[3] Piczak, KJ. Environmental sound classification with convolutional neural networks. Ieee International Workshop on Machine Learning for Signal Processing 2015. 17-20 September; Boston, USA 2015.

[4] Inoue, T; Vinayavekhin, P; Wang, S; Wood, D; Greco, N; Tachibana, R. Domestic activities classification based on CNN using shuffling and mixing data augmentation technical report. Detection and Classification of Acoustic Scenes and Events 2018.

[5] Cui, X; Goel V; Kingsbury B. Data Augmentation for Deep Neural Network Acoustic Modeling. IEEE/ACM Trans Audio, Speech, Lang Process. 23(9), 2015, 1469–77.

[6] Dekkers, G; Vuegen, L; van Waterschoot, T; Vanrumste, B; Karsmakers, P; DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics. 2018.

[7] Dekkers, G; Lauwereins, S; Thoen, B; Adhana, MW; Brouckxon, H; Bergh, B; Van Den, B; et al. The SINS database for detection of daily activities in a home environment using an acoustic sensor network. Detection and Classification of Acoustic Scenes and Events. 2017; Munich, Germany 16 November 2017; 32(0).

[8] Kingma, DP; Ba, J. Adam: A method for stochastic optimization. ICLR 2015, p. 1–15.