MULTICHANNEL ACOUSTIC SCENE CLASSIFICATION BY BLIND DEREVERBERATION, BLIND SOURCE SEPARATION, DATA AUGMENTATION, AND MODEL ENSEMBLING

Technical Report

Ryo Tanabe[1]*, Takashi Endo*[1]*, Yuki Nikaido*[1]*, Takeshi Ichige*[1]*,*
Phong Nguyen[1]*, Yohei Kawaguchi*[1]*, Koichi Hamada*[1]

[1] Hitachi, Ltd., Research and Development Group, Tokyo, Japan,
{ryo.tanabe.rw, yohei.kawaguchi.xk}@hitachi.com

ABSTRACT

Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Challenge Task 5 can be regarded as one type of multichannel acoustic scene classification. The important characteristic of the Task 5 is that a microphone array may be put at different locations between the development dataset and the evaluation dataset, so we should not exploit location-dependent spatial cues but location-independent ones to avoid overfitting. The proposed system is a combination of front-end modules based on blind signal processing and back-end modules based on machine learning. To avoid overfitting, the front-end modules employ blind dereverberation, blind source separation, etc., which use the spatial cues without machine learning. The back-end modules employ one-dimensional-convolutional-neural-network-(1DCNN)-based architectures and VGG16-based architectures for individual front-end modules, and all the 89 probability outputs are ensembled.

Index Terms— acoustic scene classification, blind dereverberation, blind source separation, convolutional neural network, model ensembling

1. INTRODUCTION

Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Challenge Task 5 is an acoustic classification task for daily activities in a home environment. The task is to estimate 9-class daily activities from the 4-channel 10-second signal obtained by a microphone array. The Task 5 is similar to the Task 1, but we can make use of spatial cues that can be extracted from the multichannel signal. Also, the Task 5 has another important characteristic that the microphone array may be put at different locations between the development dataset and the evaluation dataset. Therefore, we should not exploit location-dependent spatial cues but location-independent ones to avoid overfitting.

The proposed system is a combination of front-end modules based on blind signal processing and back-end modules based on machine learning. The front-end modules perform dereverberation, source separation, and noise reduction. For the front-end modules, many approaches based on machine learning have been proposed [1], but it can be thought that the machine-learning-based front-end modules suffers from overfitting because the development data is not sufficiently large. Also, it can be thought that external datasets such as Audio Set are not suitable for training the front-end modules because acoustic features of the Task 5 have a large gap from those of the external datasets. Therefore, the front-end modules employ blind dereverberation, blind source separation, etc., which use the spatial cues without machine learning, so overfitting is avoided. The back-end modules perform feature extraction, classification, and ensemble-based decision. In the back-end modules, log mel energy features and MFCC features are extracted for individual front-end modules, these features are given individually to 1-dimensional-convolutional-neural-network-(1DCNN)-based architectures and VGG16[2]-based architectures, and the 89 probability outputs from all the networks are ensembled. By the model ensembling, the participants aim to prevent overfitting.

This paper presents the detail of the proposed system. Experimental results for the development dataset are also shown.

2. PROPOSED SYSTEM

2.1. Whole Architecture

The whole architecture of the proposed system is shown in Fig. 1. The system consists of two disjoint parts: the first part is called ”front-end”, and the second part is called ”back-end” in this paper. The modules in the front-end part perform dereverberation, source separation, and noise reduction. The modules in the back-end part perform feature extraction, classification, and ensemble-based decision. As shown in Fig. 1, the whole system is very huge but its architecture is simple.

2.2. Front-End Modules

As explained above, we should not exploit location-dependent spatial cues but location-independent ones to avoid overfitting. The front-end modules employ blind dereverberation, blind source separation, etc., which does not use machine learning. Overfitting is therefore avoided. Moreover, multiple different front-end modules send different output signals to the back-end in parallel. It can be considered that this approach is a manner for training/test data augmentation and provides the system with robustness.

The participants assume that reverberation is an important cue about room activities, so they focus on dereverberation. For blind dereverberation (BD), an algorithm proposed by Togami et al. [3] is used for the 4-channel input signal. Togami’s BD is a multi-input-multi-output (MIMO) approach, so it outputs 4-channel dereverbed signal. In addition, the reverberation signal is obtained by subtracting the dereverbed signal from the input signal.

In addition, the participants predict that multichannel source separation extracts turn-taking features in ”social activity”, and they

use a source separation method. The system employs a blind source separation (BSS) algorithm proposed by Duong et al. [4]. The 4-channel dereverbed signal is sent to Duong's BSS. Duong's BSS is also an MIMO approach, and the number of sources is set to 2 in our system, so the 8-channel separated signal is calculated. The order of the separated sources is arbitrary, so both the original order and the inverted order are sent in parallel to back-end classifiers.

Harmonic-percussive sound separation (HPSS) [5] is also employed in our system because Han et al. [6] reports that the HPSS is suitable for acoustic classification in DCASE 2017 Task 1. The HPSS separates the (monaural) input signal into the harmonic sound and the percussive sound. The proposed system applies an non-negative matrix factorization (NMF)-based HPSS [7] similarly to the conventional work [6].

Also, inspired by the conventional work [6], the system employs a simple beamformer only by addition and subtraction. The participants expect that the simple beamformer reduces noise and extracts spatial cues. The simple beamformer requires only a very short calculation time in comparison with other methods. The simple beamformer calculates the output $\boldsymbol{y}(t) = [y_1(t), y_2(t), y_3(t), y_4(t)]^T$ from the input $\boldsymbol{x}(t) = [x_1(t), x_2(t), x_3(t), x_4(t)]^T$:

$$\boldsymbol{y}(t) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \boldsymbol{x}(t) \tag{1}$$

where y_m represents the m-th channel output, and x_m represents the m-th channel input. The first channel y_1 corresponds to the simple summation, and the other channels y_2, y_3, and y_4 correspond to beams orthogonal to each other.

From the front-end modules to the back-end modules, the system sends (1) the 4-channel input signal, (2) the 4 pairs of the reverberation signal and the dereverbed signal, (3) the 4-channel dereverbed signal, (4) the 8 pairs of the 2-source separated signals, (5) the 4 pairs of the harmonic sound and the percussive sound, and (6) the output of the simple beamformer.

2.3. Back-End Modules

The main elements of the back-end modules are classifiers for each preprocessed signal. All the classifiers have almost the same architecture. First, log mel energy features and MFCC features are extracted. The frame size is 40 ms, and the hop size is 50 %. Next, the log mel energy features are given to the 1DCNN-based baseline network [8], the pre-trained VGG16 [2] connected with 3 dense layers, the pre-trained VGG16 connected with a support vector machine (SVM), and the fine-tuned VGG16 connected with 3 dense layers (1024-128-32 units). There are many pre-trained open models, so we have compared the performance of them experimentally. We have confirmed that the VGG16 has the most suitable performance for this task. The number of mel filters is set to 50 and 128 for the 1DCNNs and the VGG16s, respectively. For the VGG16s receiving the raw input signal, the signal consisting of the 3 copied channels is input because the VGG16 can receive only 3-channel color images. Also for the VGG16s receiving the pair of signals, the signal converted to a 3-channel combination, i.e., the signals from Togami's BD is converted to $(\text{dereverbed}, \text{dereverbed}, \text{reverberation})$, the signals from Duong's BSS is converted to $(\text{source1}, \text{source1}, \text{source2})$, and the signals from the HPSS is converted to $(\text{harmonic}, \text{harmonic}, \text{percussive})$. The VGG16s receiving the simple-beamformed signal use the 3-channel consisting of y_1, y_2, and y_3. The VGG16 is not used for the sole dereverbed signal or the Duong's BSS separated signal because training for it has not finished until the deadline. The MFCC features are used by the 1DCNN-based baseline network. Then, all the classifiers output the probabilities that the 10-second input signal belongs to each class. Each classifier is common to all the 4 microphones, and the classifier is also trained by all the 4 microphones. It can be considered that this approach is a manner for training/test data augmentation and provides the system with robustness.

In the late fusion module, the 89 output probabilities from all the classifiers are ensembled. One is selected from four methods: the first method is probability averaging, the second is the random forest classifier, the third is the SVM classifier, and the fourth is "F1-score-weighted probability averaging". Both the random forest classifier and the SVM classifier are trained by the pairs of the output probabilities from all the classifiers and the supervision labels. In "F1-score-weighted probability averaging", the probabilities of each classifier are weighted by the square of the worst class-wise F1 score for the classifier, and the final scores are calculated by averaging the weighted probabilities over all the classifiers. Overfitting is prevented by these ensemble approaches.

3. EXPERIMENT

3.1. DCASE 2018 Challenge Task 5 Dataset

The DCASE 2018 Task 5 dataset, which is a derivative of the SINS Database [9], includes 9 scenes which are absence, cooking, dishwashing, eating, other, social activity, vacuum cleaning, watching TV, working. The development dataset consists of total 268 sessions, which include total 72984 segments of 10 seconds. The development dataset was recorded at 16 kHz with 16 bit per sample by using the microphone arrays at 4 different locations.

3.2. Results

Table 1 shows the F1 scores for each system and each class. The classification performances of the proposed system are higher than that of the baseline system.

4. SUBMISSION

The participants submitted the results the four late-fusion systems yielded: the first is ensembling by probability averaging (submission 1), the second is ensembling by random forest (submission 2), the third is ensembling by SVM (submission 3), and the fourth is ensembling by F1-score-weighted probability averaging (submission 4). In submission, the classifiers were trained for each fold and used for ensembling the fold-wise classifiers.

5. CONCLUSION

For DCASE 2018 Task 5, we proposed a system of multichannel acoustic scene classification. The system is a combination of front-end modules based on blind signal processing and back-end modules based on machine learning. As a result, the 89 probability outputs from all the back-end classifiers are ensembled. The participants believe that the architecture of the system is robust to overfitting. Evaluation results for the development dataset indicate that the proposed system improves the classification performance.

Table 1: F1 scores for the development dataset

Class	Baseline	Proposed (mean prob.)	Proposed (random forest)	Proposed (SVM)	Proposed (mean F1-weighed prob.)
Absence	85.41	90.63	87.91	87.36	90.49
Cooking	95.14	96.28	96.54	96.58	96.37
Dishwashing	76.73	83.78	86.43	86.82	84.54
Eating	83.64	92.92	93.93	94.40	93.15
Other	44.76	60.29	62.29	64.76	61.01
Social activity	93.92	96.23	96.26	95.61	95.52
Vacuum cleaning	99.31	100.00	100.00	100.00	100.00
Watching TV	99.59	99.59	99.70	99.54	99.38
Working	82.03	88.06	87.34	87.26	88.07
Average	84.50	89.75	90.04	90.26	89.84

6. REFERENCES

[1] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, June 2015.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv:1409.1556*, 2015.

[3] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1369–1380, July 2013.

[4] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.

[5] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. 16th European Signal Processing Conference (EUSIPCO)*, Aug. 2008, pp. 1–4.

[6] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov. 2017.

[7] J. Park, J. Shin, and K. Lee, "Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1061–1074, May 2017.

[8] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 challenge - task 5: Monitoring of domestic activities based on multi-channel acoustics," in *arXiv:1807.11246*, 2018.

[9] G. Dekkers, S. Lauwereins, B. Thoen, M. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov. 2017, pp. 32–36.

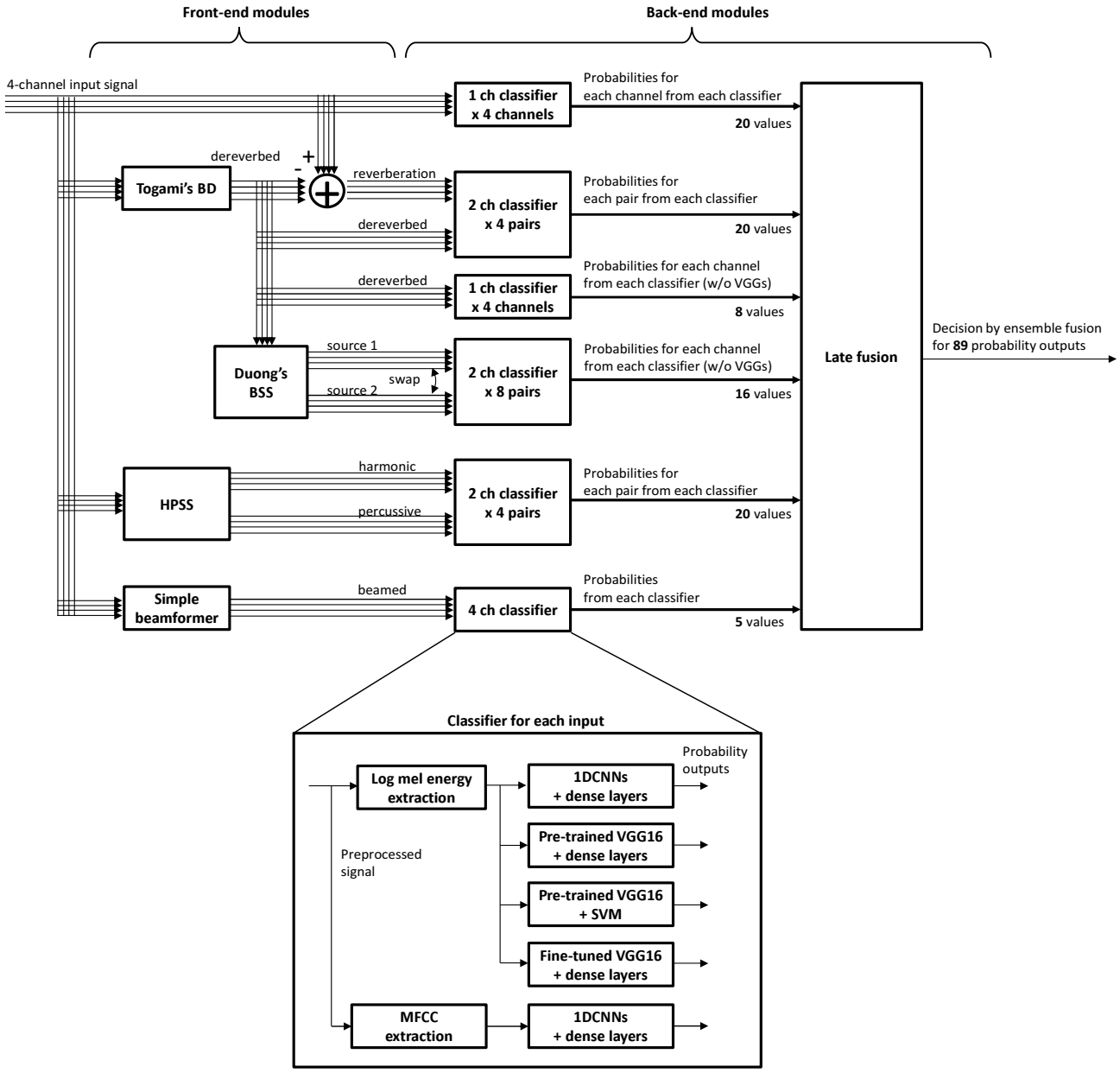

Figure 1: Layout of the whole system