Location-Independent Multi-Channel Acoustic Scene Classification Using Blind Dereverberation, Blind Source Separation, and Model Ensemble

Ryo Tanabe, Takashi Endo, Yuki Nikaido, Kenji Ichige, Nguyen Phong, Yohei Kawaguchi, and Koichi Hamada
Research and Development Group, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
Email: ryo.tanabe.rw@hitachi.com

***Abstract*—This paper presents a location-independent multi-channel acoustic scene classification (ASC) system that avoids spatial overfitting. Generally, ASC suffers from noise and reverberation in real environments. In addition, the ASC performance is decreased by overfitting a dataset, which is the result of learning from acoustic transfer functions enclosed in the dataset. To resolve these problems, we present a location-independent multi-channel ASC system using blind dereverberation, blind sound source separation, pre-trained model-based classifiers, and model ensemble. Experimental results on the DCASE 2018 Task 5 dataset indicate that the proposed system, with an F1 score of 88.4%, outperforms the baseline system. Also, the results indicate that although no one specific function improves the performance dramatically, all functions complement each other through the model ensemble.**

***Index Terms*—acoustic scene classification, blind dereverberation, blind source separation, pretrained model, model ensemble**

I. Introduction

Multi-channel acoustic scene classification (ASC) is a technology that classifies an audio signal into a predefined class characterizing the environment in which it was recorded. It is expected to be used for machine health monitoring and other applications. In real environments, ASC suffers from noise and reverberation, whereas multi-channel ASC can solve this problem by using a microphone array.

The purpose of this paper is to propose a "location-independent" multi-channel ASC. Location-independence means robustness to the change of acoustic transfer functions (ATFs). In real scenarios, ATFs differ between a training and evaluation datasets because locations of sources and microphones may change. If the ATFs enclosed in the training dataset are learned as they are, spatial overfitting will occur, i.e., the performance for the evaluation dataset tends to be severely degraded due to the change of cues for ASC. In addition to the spatial overfitting, overfitting is caused by other factors [1][2]. Robustness is required against both spatial overfitting and overfitting caused by other factors.

We present a novel location-independent multi-channel ASC system consisting of pre-processing, classifiers, and model ensemble. The core idea of this paper is that various pre-processing and classifiers each have strengths and weaknesses, i.e., Harmonic-percussive sound separation (HPSS) [3][4] is suitable for ASC but may overfit. Pre-processing uses blind dereverberation (BD) [5], blind sound source separation (BSS) such as Duong's BSS [6] and HPSS and, a beamformer [7]. On the basis of these pre-processing algorithms, the proposed system solves the problem of noise and reverberation and avoids spatial overfitting. In addition, classifiers use a pre-trained convolutional deep neural network (CNN) model, VGG16 [8], to classify images. By using VGG16, the proposed system avoids both spatial overfitting and overfitting caused by other factors. Moreover, the model ensemble is the core method of the proposed ASC system to improve the robustness by fusing the strengths of pre-processing and classifiers.

In the experiment section, we verify the effects of the proposed system on ASC performance using the DCASE 2018 Task 5 [9] dataset [10]. The dataset was recorded in an indoor environment, and the audio signals in the dataset include a lot of noise and reverberation. Experimental results indicate that the proposed system, with an F1 score of 88.4%, works well when locations of sources and microphones may change, outperforms the baseline system, and tied for first place in DCASE 2018 Task 5. In addition, results indicated that no one specific function implemented as pre-processing or a classifier improves the performance dramatically, but all functions complement each other through the model ensemble.

II. RELATION TO PRIOR WORK

To achieve location-independent multi-channel ASC, the proposed system has the unique point that it uses dereverberation such as Togami's BD [5]. In ASC, many works [7][11][12] are related to the proposed system, but dereverberation is not used in the previous works. Han and Park [7] achieved second place in DCASE 2017 Task 1 using binaural audio, HPSS, background subtraction, and model ensemble. Sakashita and Aono [11] used binaural audio, monaural audio, HPSS, mixup-based [13] data augmentation, and model ensemble and took first place in DCASE 2018 Task 1. Inoue et al. [12] aimed to solve the same problem as this paper by using data augmentation, and their system tied with the proposed system for first place in DCASE 2018 Task 5. In addition, although there are several works [14][15] on ASC based on an image classification model, there is no work on ASC based on a combination of dereverberation and an image classification model such as VGG16. Mun et al. [14] reported the effect of the transfer learning for a model pre-trained by ImageNet dataset [16]. Hwiyong et al. [15] suggested that VGG16 [8] could be used for ASC.

The main contribution of this paper is examining the effectiveness of the proposed system architecture. In the previous work [17], we verified only the feasibility of the proposed system, so this paper studies its details and effectiveness.

III. DCASE 2018 Task 5 dataset

We use the DCASE 2018 Task 5 dataset for training and evaluation. The dataset contains real life audio recorded in an indoor environment using a 4-channel microphone array. It has nine audio classes: "Absence," "Cooking," "Dishwashing," "Eating," "Other," "Social activity," "Vacuum cleaning," "Watching TV,“ and "Working." As the audio signals include a lot of reverberation and various noises, the system needs to use dereverberation and sound source separation for robustness. The microphone array's positions in the training dataset

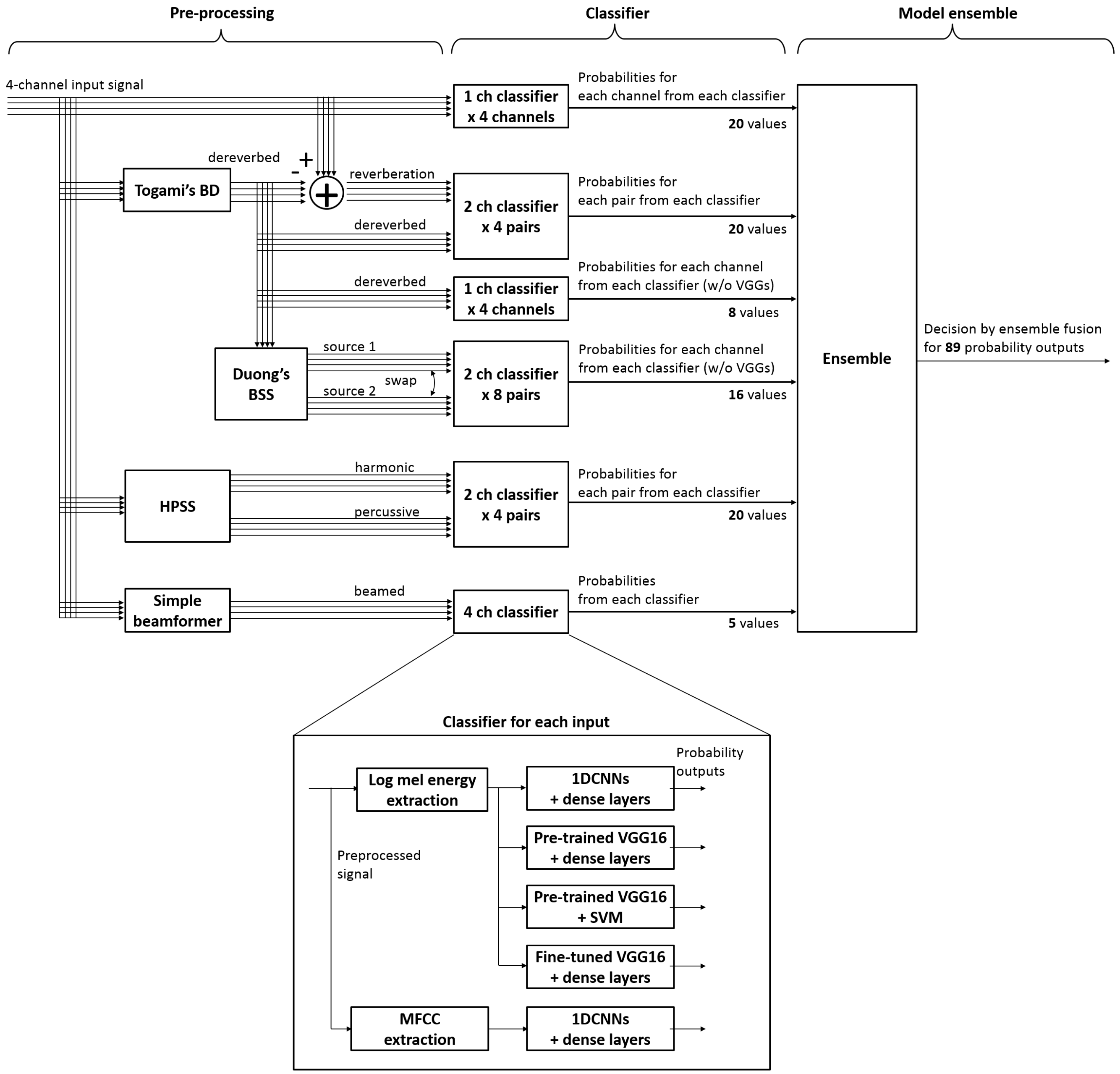
Pre-processing
Classifier
Model ensemble
4-channel input signal
1 ch classifier x 4 channels
Probabilities for each channel from each classifier
20 values
dereverbed
+
-
Togami's BD
reverberation
2 ch classifier x 4 pairs
Probabilities for each pair from each classifier
20 values
dereverbed
dereverbed
1 ch classifier x 4 channels
Probabilities for each channel from each classifier (w/o VGGs)
8 values
source 1
Duong's BSS
swap
source 2
2 ch classifier x 8 pairs
Probabilities for each channel from each classifier (w/o VGGs)
16 values
Ensemble
Decision by ensemble fusion for 89 probability outputs
harmonic
HPSS
percussive
2 ch classifier x 4 pairs
Probabilities for each pair from each classifier
20 values
beamed
Simple beamformer
4 ch classifier
Probabilities from each classifier
5 values
Classifier for each input
Probability outputs
Log mel energy extraction
1DCNNs + dense layers
Pre-trained VGG16 + dense layers
Pre-trained VGG16 + SVM
Fine-tuned VGG16 + dense layers
Preprocessed signal
MFCC extraction
1DCNNs + dense layers

Fig. 1. Layout of the proposed system

and evaluation dataset are different. Therefore, the dataset is suitable for evaluating the proposed system.

IV. LOCATION-INDEPENDENT MULTI-CHANNEL ACOUSTIC SCENE CLASSIFICATION

In this section, we first present an overview of our proposed location-independent multi-channel ASC system and then describe it in more detail.

A. Overview

The proposed system is structured into pre-processing, classifiers, and model ensemble. Fig. 1 shows the system overview.

First, in pre-processing, the proposed system uses BD and BSS. Considering that the source position of the audio can differ, as mentioned in Section III, the system uses a blind algorithm for location independence. In addition, machine learning is not used to avoid overfitting. After pre-processing, various processed audio signals are input to the classifiers.

Second, in the classifiers, a CNN is used. As previously mentioned, when using machine learning, overfitting can cause ASC performance to worsen. To avoid this, we use an open-source pre-trained model. By using this model, the proposed system reduces learning of the dataset to a minimum. The proposed system uses a pre-trained two-dimensional CNN (2DCNN) and a one-dimensional CNN (1DCNN) trained by the dataset. The various classifiers predict the audio classes from the pre-processed signal in parallel, and the output is used for the model ensemble in a later step.

Finally, in the model ensemble, the audio class is determined from a lot of prediction results generated by many pre-processing functions and various classifiers outputs. Since the contributions to the final decision for each classifier are decreased and the various predictions are combined, the proposed system avoids overfitting.

B. Pre-process

For BD, the proposed system uses Togami's BD algorithm [5], which is a multi-input-multi-output (MIMO) method. In the system, a 4-channel audio signal $S_i^{in} (i = 1, ..., 4)$ is input, and a 4-channel dereverbed audio signal S_i^d is output. Furthermore, reverberation S^r is generated from S_i^{in} and S_i^d.

In addition, the system uses Duong's BSS method [6]. Duong's BSS is also a MIMO method, and an audio signal is separated from S_i^{in} into two audio signals (S_i^{sep1} and S_i^{sep2}) per channel. Here, S_i^{sep1} and S_i^{sep2} are output in an arbitrary order; therefore, the proposed system needs to make these signals swappable in the classifier. The signals from Duong's BSS are used for solving in only the 1DCNN.

The proposed system uses HPSS [3] because it is reported to be suitable for ASC. HPSS separates an audio signal S_i^{in} into a harmonic audio signal S_i^{har} and percussive audio signal S_i^{per}. Non-negative matrix factorization based HPSS [4] is used.

The proposed system implements simple beamforming (Simple BF) only by addition and subtraction because Han and Park reported that a similar method is suitable for ASC [7]. Simple BF only needs a very short calculation time and calculates the output $\boldsymbol{S^{BF}} = [S_1^{BF}, S_2^{BF}, S_3^{BF}, S_4^{BF}]^T$ using $\boldsymbol{x} = [S_1^{in}, S_2^{in}, S_3^{in}, S_4^{in}]^T$:

$$\boldsymbol{S^{BF}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \boldsymbol{x} \tag{1}$$

where, S_1^{BF} is the result of the simple summation, and S_2^{BF}, S_3^{BF}, and S_4^{BF} are the output of beams orthogonal to each other.

C. Classifier

As mentioned in Section IV-A, the proposed system uses a 1DCNN and a 2DCNN (VGG16 [8]). Many pre-trained models [18] [19] are available, so we compared their performances experimentally. We confirmed that VGG16 has the most suitable performance for this task. Almost all the classifiers have the same architecture. First, log mel energy features and mel-frequency cepstral coefficient (MFCC) features are extracted. The frame size is 40 ms, and the hop size is 50%. Next, these features are sent to the classifiers. MFCC features are sent to a baseline network [9]. The log mel energy features are sent to the 1DCNN-based baseline network, the pre-trained VGG16 connected with three dense layers (1024-128-32 units), the pre-trained VGG16 connected with a support vector machine (SVM), and the fine-tuned VGG16 connected with three dense layers. The numbers of mel filters are set to 50 and 128 for the 1DCNNs and VGG16s, respectively.

Because the VGG16 can receive only 3-channel color images from the raw input signal, a signal consisting of the three copied channels is input. On the other hand, for the VGG16 receiving a pair of signals, the signals are converted into a 3-channel combination, e.g., the signals from Togami's BD and the HPSS are converted into (S_i^d, S_i^d, S_i^r) and (S_i^{har}, S_i^{har}, S_i^{per}), respectively. In addition, for the VGG16s receiving the simple-beamformed signal, the three channels consist of ($S_1^{BF}, S_2^{BF}, S_3^{BF}$).

D. Model ensemble

There are four versions of the proposed system for model ensemble: probability averaging, random forest classifier, SVM classifier, and "F1 score-weighted probability averaging." Both the random forest and SVM classifiers are trained by the pairs of the predicted 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. The 89 output probabilities from all the classifiers are combined.

In Section V, we compare the F1 scores of the four versions and verify the effect of the proposed system.

V. EXPERIMENT

We verified the effective of the proposed system on ASC performance using the DCASE 2018 Task 5 dataset. In this section, we first describe the effectiveness of pre-processing and classifiers on the training dataset and then describe the performance of the proposed system.

A. Effectiveness of pre-processing and classifier

In this section, we evaluate the effect of pre-processing and classifiers in the proposed system.

For the evaluation, we compared the performances when each pre-processing function and classifier was removed. The reason for this is that evaluating the overall performance using only one pre-processing function or classifier will only indicate its comparative superiority and is insufficient for evaluating its contribution to increasing performance. For example, when a classifier is removed and the performance then degrades, it means the classifier contributes to increasing performance. In addition, when a function is removed and the performance does not change, it means that the function

TABLE I

F1 SCORES FOR THE DCASE 2018 DATASET. "ALL OUTPUT" REFERS TO THE RESULT OF MODEL ENSEMBLE USING ALL OUTPUT PROBABILITIES OF THE CLASSIFIERS. THE SECOND SECTION REFERS TO THE F1 SCORES WHEN EACH PRE-PROCESS IS REMOVED OR USED ONLY. THE THIRD SECTION REFERS TO THE F1 SCORES WHEN EACH CLASSIFIER IS REMOVED OR USED ONLY. EACH METHOD IS EVALUATED USING THE TRAINING DATA (TRAIN.) AND EVALUATION DATA (EVAL.).

Functions			Remove each proc.	Only each proc.
Pre-process	All output	train.	89.75	
		eval.	89.12	
	Input signal	train.	89.78	88.46
		eval.	88.84	88.48 (+0.02)
	Togami's BD	train.	89.67	89.06
		eval.	89.13	88.26 (-0.80)
	Duong's BSS	train.	89.75	85.66
		eval.	89.13	85.07 (-0.59)
	HPSS	train.	89.47	90.08
		eval.	89.17	88.26 (-1.82)
	Simple BF	train.	89.67	88.85
		eval.	88.93	88.27 (-0.58)
Classifier	VGG16 + fc-layer	train.	89.36	89.11
		eval.	88.97	87.66 (-1.45)
	VGG16 + SVM	train.	89.05	87.51
		eval.	87.94	86.67 (-0.84)
	1DCNN	train.	89.15	87.23
		eval.	88.24	86.30 (-0.93)

TABLE II

F1 scores for DCASE 2018 dataset. “Baseline” gives the results for DCASE 2018 baseline system, “Proposed (mean prob.)” for an ensemble using mean of predicted probability, “Proposed (RF)” for a random forest (RF) as an ensemble, “Proposed (SVM)” for a support vector machine (SVM) as an ensemble, and “Proposed (F1-weighed mean prob.)” for an ensemble using weighted mean of predicted probability based on F1 score.

Class	Baseline		Proposed		Proposed		Proposed		Proposed	
			(mean prob.)		(F1-weighted mean prob.)		(RF)		(SVM)	
	dev.	eval.	dev.	eval.	dev.	eval.	dev.	eval.	dev.	eval.
Absence	85.4	87.7	90.6	91.6	90.4	91.3	87.9	59.1	87.3	86.1
Cooking	95.1	93.0	96.2	97.0	96.3	97.0	96.5	96.1	96.5	95.8
Dishwashing	76.7	77.2	83.7	83.0	84.5	83.0	86.4	81.5	86.8	81.6
Eating	83.6	81.2	92.9	84.2	93.1	84.1	93.9	85.7	94.4	85.2
Other	44.7	35.0	60.2	57.7	61.0	58.3	62.2	53.7	64.7	54.6
Social activity	93.9	96.6	96.2	98.2	95.5	98.2	96.2	97.7	95.6	95.9
Vacuum cleaning	99.3	95.8	100.0	97.7	100.0	97.7	100.0	97.7	100.0	96.7
Watching TV	99.5	99.9	99.5	100.0	99.3	100.0	99.7	100.0	99.5	100.0
Working	82.0	81.4	88.0	86.1	88.0	85.8	87.3	68.6	87.2	81.3
Average	84.5	83.1	89.7	88.4	89.8	88.4	90.0	82.2	90.2	86.3

contributes to increasing the complement and the robustness but not the numerical performance. Table I shows the performances of the proposed system with each pre-processing or classifier removed and the performances using only each pre-processing function or classifier.

For pre-processing, in the upper section of Table I, the performance when HPSS is removed (89.47%) is the most different from that of “all output” (89.75%) and the performance of only HPSS is the highest (90.08%) for the training dataset. However, the performance when HPSS is removed (89.17%) is almost the same as that of “all output” (89.12%) and the performance of only HPSS worsens (-1.82%) for the evaluation dataset. This means that HPSS improves the performance in the training [but is overfitting to the training dataset. In addition, in the evaluation, the performance when Togami’s BD is removed is 89.13%. The difference between training and evaluation is smaller in the only-Togami’s BD case (-0.80%) than in the only-HPSS case (-1.82%). This indicates Togami’s BD avoids overfitting and makes an average contribution to improve the performance in the proposed system. On the other hand, the performance when Simple BF is removed is 88.93% and the performance of only Simple BF slightly decreases (-0.58%) in the evaluation dataset. Therefore, Simple BF avoids overfitting and contributes to improving the performance in the proposed system. Additionally, the performance of only “input signal” barely changes between training and evaluation (+0.02%), and the performance with it is removed (88.84%) is less than that of “all output” performance (89.75%). This means “input signal” is not overfitting and contributes to improving the F1-score. The performance when Duong’s BSS is removed does not differ from that of “all output” performance (89.75%) even though the performance of only Duong’s BSS is the lowest (85.07%). This is because Duong’s BSS is used in only the 1DCNN, which has little effect on classification.

For the classifiers, in the lower section of Table I, the performance when VGG16 + SVM’s is removed (89.05%) is the most different from that of “all output” (89.75%) for the training dataset. However, the other classifiers also tend to have larger differences from “all output” than the pre-processing functions do.

These results indicate that no one specific function improves the performance dramatically, but all functions complement each other through the model ensemble. Also, VGG16 is suitable for location-independent multi-channel ASC even if it uses a pre-trained model for image classification.

B. *Performance of proposed system*

To evaluate the location-independence, we evaluate the proposed system with the dataset which the position of the microphone array differs between the training and evaluation dataset. Table II shows the experimental results of the four versions of our proposed system and the DCASE 2018 baseline-system on the dataset.[1] The baseline-system is based on a 1DCNN and is a state-of-the-art ASC system. For the training dataset, each proposed system outperforms the “Baseline,” 84% F1 score and achieves an F1 score around 90%. All proposed systems achieved a 100% F1 score for “Vacuum cleaning’’ in the evaluation. In addition, all proposed systems had F1 scores over 7% higher than that of “Baseline,” in “Dishwashing,” “Eating,” “Other,” and “Working,” with “Other” marking the largest improvement with F1 scores over 15% higher. The “Other” class is unique in that it consists of miscellaneous data. Data in “Other” is more difficult to classify than methodical data in other classes (“Vacuum cleaning,” “Cooking,” etc.). This improvement suggests that the proposed system is robust to unknown data found in the training dataset.

Results on the evaluation dataset suggest the effectiveness of the proposed location-independent multi-channel ASC system because the F1 scores for the training and evaluation datasets are only 1.5% different on average in the cases of “mean prob.” and “F1-weighted mean prob.” These results indicate that the proposed systems have a high location-independence because the position of the microphone array differs between the training and evaluation datasets. The proposed system (88.4%) tied for first place in DCASE 2018 Task 5. This challenge result indicates that the proposed system outperforms the other state-of-the-art systems. The performances of “RF” and “SVM,” however, were about 7.8% and 3.9% lower, respectively. The results for “RF” indicate a large difference in performance between the training and evaluation datasets. Specifically, the “RF” performances for “Absence,” “Other,” and “Working” were by respectively about 20%, 10%, and 20% lower in the evaluation dataset than in the training dataset. The results for “SVM” indicate that the performances for “Other” and “Working” were also about 10% lower in the evaluation dataset than in the training dataset.

[1] There is some differences between Tables I and II because the evaluation method is different. Table I shows the results using all the evaluation dataset. Table II shows the results using only the “location unknown microphone” dataset, which is published by the task organizer. The authors asked the task organizer about the property of the dataset recording and have not received a reply at the time of writing.

The results suggest that the machine learning-based ensemble (e.g., "RF" and "SVM") overfitted to the dataset and the non-machine-learning based ensemble avoided overfitting.

VI. CONCLUSION

We presented a location-independent multi-channel acoustic scene classification (ASC) system that avoids spatial overfitting generally caused by learning from spatial information enclosed in dataset. The proposed ASC system uses blind dereverberation, blind sound source separation, pre-trained model-based classifiers, and a model ensemble. Experimental results on the DCASE 2018 Task 5 dataset indicate that the proposed system, with an F1 score of 88.4%, outperforms a baseline system. From the experimental results, location-independent multi-channel ASC based on non-machine-learning ensemble is effective for robust ASC. In addition, results indicated that no one specific function improves the performance dramatically, but all functions complement each other through the model ensemble.

REFERENCES

[1] Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 2, pp. 379–393, 2018.

[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Acoustic scene classification: an overview of DCASE 2017 challenge entries," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 411–415.

[3] Nobutaka Ono, Kenichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka, and Shigeki Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–4.

[4] Jeongsoo Park, Jaeyoung Shin, and Kyogu 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, 2017.

[5] Masahito Togami, Yohei Kawaguchi, Ryu Takeda, Yasunari Obuchi, and Nobuo Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1369–1380, 2013.

[6] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] Yoonchang Han and Jeongsoo Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," Tech. Rep., DCASE2017 Challenge, September 2017.

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] Gert Dekkers, Lode Vuegen, Toon van Waterschoot, Bart Vanrumste, and Peter Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," *arXiv preprint arXiv:1807.11246*, 2018.

[10] Gert Dekkers, Steven Lauwereins, Bart Thoen, Mulu Weldegebreal Adhana, Henk Brouckxon, Bertold Van den Bergh, Toon van Waterschoot, Bart Vanrumste, Marian Verhelst, and Peter Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Detection and Classification of Acoustic Scenes and Events 2017*, 2017.

[11] Yuma Sakashita and Masaki Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," Tech. Rep., DCASE2018 Challenge, September 2018.

[12] Tadanobu Inoue, Phongtharin Vinayavekhin, Shiqiang Wang, David Wood, Nancy Greco, and Ryuki Tachibana, "Domestic activities classification based on CNN using shuffling and mixing data augmentation," Tech. Rep., DCASE2018 Challenge, September 2018.

[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[14] Seongkyu Mun, Suwon Shon, Wooil Kim, David K Han, and Hanseok Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 796–800.

[15] Choi Hwiyong, Lee Seungjun, Yang Haesang, and Seong Woojae, "Classification of noise between floors in a building using pre-trained deep convolutional neural networks," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 535–539.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[17] Ryo Tanabe, Takashi Endo, Yuki Nikaido, Kenji Ichige, Phong Nguyen, Yohei Kawaguchi, and Koichi Hamada, "Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling," Tech. Rep., DCASE2018 Challenge, September 2018.

[18] François Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint*, pp. 1610–02357, 2017.

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.