Closed-Form Pre-Training for Small-Sample Environmental Sound Recognition

Nakamasa Inoue*, Keita Goto*
Tokyo Institute of Technology
E-mail: inoue@c.titech.ac.jp

***Abstract*—This paper presents a framework for pre-training neural networks, namely *closed-form pre-training*, and we apply it to small-sample environmental sound recognition. Our main idea is to pre-train neural networks on a dataset automatically generated by some formulas, without any prior real-world recordings or manual annotation. Specifically, the proposed framework consists of two steps. First, an audio classification dataset is generated. Here, we propose three types of dataset definitions using colored noise and its extensions. Second, a network is pre-trained on the generated dataset. The obtained pre-trained network is particularly effective for fine-tuning with few examples because it helps optimization methods avoid falling into a premature local optimal solution. In experiments, we demonstrate the effectiveness of the proposed framework for small-sample environmental sound recognition on three datasets: ESC-10/50, and UrbanSound8K. We obtained performance improvement on all datasets with a small number of training samples.**

I. Introduction

Environmental sound recognition is an interesting research topic having various applications to search, surveillance, adaptive speech recognition, and robotics. Thanks to the development of learning techniques, recent research has made notable progress in understanding environmental sounds. In particular, some studies have shown that statistical approaches using deep neural networks are effective.

To train neural networks, most optimization methods require a set of training audio samples labeled with sound categories. For example, to train convolutional neural networks (CNNs), such as SoundNet [1] and EnvNet [2], manually labeled datasets, such as ESC-50 [3] and UrbanSound8K [4], are often used. After training, these networks can classify audio samples into up to 100 environmental sound categories. However, they are often sensitive to differences in recording conditions. This is because, compared with automatic speech recognition, the target environmental sounds are less structured. In fact, if we compare human speech and environmental sounds such as airplane engine sounds, the latter are more similar to recording noise than the former.

To avoid the sensitivity problem in practice, networks are typically trained using recordings collected from the same location where a sound recognition system will be used. However, since the recording and manual labeling are often costly, it is not always realistic to prepare enough audio samples to train networks. Therefore, few-example environmental sound recognition, which aims to train networks from a given set of few audio samples per category, is needed.

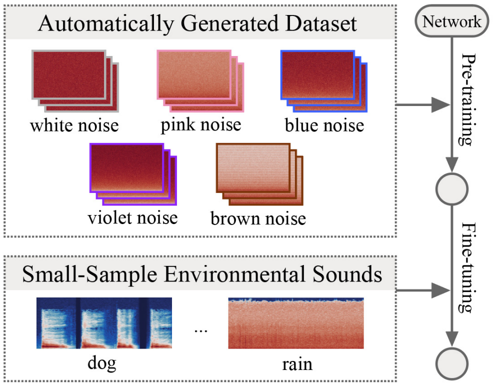

Fig. 1. Overview of closed-form pre-training for small-sample environmental sound recognition. A network is first pre-trained on an automatically generated dataset, and then fine-tuned using a few recording examples. The Color5 dataset with 5 classes of colored noise is illustrated.

To effectively utilize a few examples, a promising approach is to apply an adaptation technique. For example, fine-tuning is the most popular technique, in which network parameters are first pre-trained in a source domain, and then fine-tuned in a target domain. In general, increasing the size of the dataset for pre-training improves the classification performance in the target domain. However, collecting a large number of audio samples with valid copyrights and licenses is not easy in practice.

This paper explores a framework for pre-training, namely *closed-form pre-training*, which aims to pre-train networks with a dataset automatically generated without using real-world sound recordings. Specifically, the proposed framework consists of two steps. First, an audio classification dataset is generated. Here, we propose three types of dataset definitions using colored noise and its extensions. Second, a network is pre-trained on the generated dataset. Figure 1 provides an overview of few-example environmental sound recognition using the proposed framework. In experiments, we evaluated the proposed framework on the ESC-10, ESC-50, and UrbanSound8K datasets, and showed the effectiveness of the proposed method for few-example environmental sound recognition. In summary, our main contributions are threefold:

1) Development of a new closed-form pre-training framework using artificially generated audio noise samples.
2) Definition of three noise datasets using colored noise and its extensions.
3) Exhaustive experiments with few-example environmental

[*] equal contribution

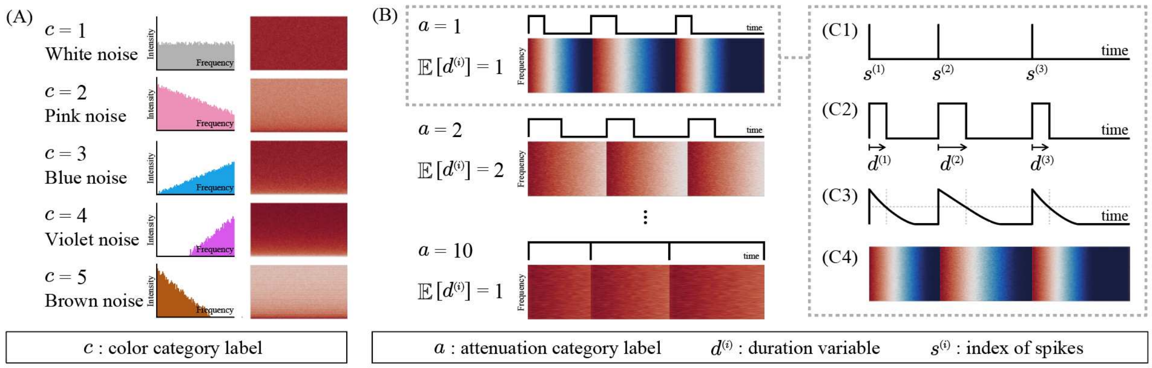

Fig. 2. Visualization of dataset generation. (A) Five color categories for the colored noise dataset. The frequency distribution of each category and generated audio sample spectrum are shown. (B) Ten attenuation categories for the Poisson colored noise dataset. Detailed steps are visualized from (C1) to (C4).

sound recognition on publicly available datasets.

The remainder of the paper is organized as follows. Sec. 2 reviews related work on environmental sound recognition. Sec. 3 describes our method and the dataset definitions. Sec. 4 presents and discusses the experimental results. Finally, conclusions are offered in Sec. 5.

II. Related Work

A. Environmental Sound Recognition

Environmental sound recognition is a task to classify audio samples into pre-defined sound categories. Over the past 10 years, statistical signal processing approaches have led to great success in this task. For example, probabilistic models, such as Gaussian mixture models and hidden Markov models, effectively estimate the distributions of audio features, such as mel frequency cepstral coefficients. Recent studies have moved to focus on end-to-end training of neural networks.

Dai et al. [5] proposed a deep CNN that accepts raw audio data as input. Sailor et al. [6] proposed a convolutional restricted Boltzmann machine to learn filter-bank features. Temporal modeling with recurrent neural networks [7], [8], and attention mechanisms [9], [10] are also known to be effective. Tokozume et al. [2], [11] reported that EnvNet with a swapping layer further improved the classification accuracy.

For evaluating these networks, researchers proposed datasets compiled from environmental recordings. Piczak provided the environmental sound classification dataset (ESC-10/50), which consists of 2,000 audio samples of sound categories such as *sea waves* and *chirping birds*. Salamon et al. constructed UrbanSound8k [4] with 8,000+ urban recordings. Rakotomamonjy et al. [12] provided the LITIS Rouen dataset for acoustic scene classification. More challenging audio datasets are provided in the DCASE workshop series [13], [14].

B. Training with Few Examples

Training with few examples is a relatively new topic in the field of audio and speech processing. For sound recognition, including audio event detection, previous studies have found that adaptation and fine-tuning techniques [15], [16], [17] help improve the recognition accuracy. Data augmentation techniques are also known to be effective for detecting rare sound events [18]. For speech recognition, transfer learning techniques from a high-resource domain to low-resource domains are proposed in [19], [20]. A model pre-trained on a large-scale English corpus is often adapted to low-resource languages.

In these studies, pre-training on a large-scale dataset plays an important role. However, collecting such a dataset is costly, and its use is often limited to non-commercial research and education uses. To complement the previous work, this study explored a new technique for pre-training that does not require any up-front costs for sample collection or manual annotation.

III. Proposed Method

This section presents the proposed pre-training framework, namely closed-form pre-training, for few-example environmental sound recognition. Our main idea is to pre-train neural networks by using datasets artificially generated with formulas. Specifically, we propose three types of datasets based on colored noise.

A. Notation and Overview

Let $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ be a training set for environmental sound recognition, where x_i is an audio sample and y_i is a sound category label. Here, we assume that few examples are given for each category. Our goal is to train a neural network $\mathcal{N}_\theta$ that predicts sound categories for new test samples, where θ is a set of network parameters.

Recent neural networks [2], [6] have a large number of parameters, and their performance improves with increased network size. For these networks, training with few examples is a challenging task because parameter optimization methods often fall into a premature local optimal solution.

To avoid this problem, the proposed framework pre-trains neural networks on an artificially generated dataset. Specifically, it consists of two steps: dataset generation and parameter estimation. In the first step, an audio classification dataset $\mathcal{E} = \{(\epsilon_j, w_j)\}_{j=1}^{M}$ is generated, where ϵ_j is an audio sample and w_j is a category label. This dataset is independent from $\mathcal{D}_{train}$, but the input format, including sampling rate, should be the same. In the second step, the network $\mathcal{N}_\theta$ is pre-trained on the generated dataset $\mathcal{E}$. The obtained pre-trained network is then fine-tuned on $\mathcal{D}_{train}$ for evaluation. The following subsections describe the details of each step.

B. Dataset Generation

Here, we propose three types of dataset definition for constructing $\mathcal{E}$. Note that none of them require any recording or annotation costs.

1) Colored Noise Dataset: The first dataset is a colored noise dataset consisting of five types of colored noise: white noise, pink noise, blue noise, violet noise, and brown noise. Pre-training on this dataset predisposes networks to focus on differences in frequency distribution. Pairs of an audio sample and a category label (ϵ_j, w_j) for $j = 1, 2, \cdots, M$ are generated by the following three steps:

1. Apply uniform random sampling to select a color category label c_j from $C = \{1, 2, 3, 4, 5\}$, where 1 to 5 denote white, pink, blue, violet, and brown, respectively. This label is identical to the label for pre-training, that is, $w_j = c_j$. The number of categories is $|C| = 5$.
2. Generate a white-noise source z_j with a duration of 5 seconds. Note that the duration should be longer than the input length of the network. In general, 5 seconds is long enough.
3. Apply the coloring filter F_{c_j} to z_j to obtain a colored noise sample by

$$\epsilon_j = F_{c_j}(z_j). \tag{1}$$

See the appendix for the filter definitions for each color.

Figure 2A visualizes the five categories. As can be seen, the frequency distribution is different from category to category. Therefore, pre-training on this dataset predisposes networks to focus on the frequency distribution of audio samples.

2) Poisson Colored Noise Dataset: To introduce time-domain variability into the dataset definition, a second dataset, namely the Poisson colored noise dataset, generates non-static noise. The noise occurrence follows the Poisson process shown in Figure 2B. The dataset generation procedure is as follows:

1. Randomly sample a color category label c_j and an attenuation category label a_j from $C = \{1, 2, \cdots, 5\}$ and $A = \{1, 2, \cdots, 10\}$, respectively. The label for pre-training is defined by $w_j = (c_j, a_j)$. This means that the number of categories is $|C \times A| = 5 \times 10 = 50$.
2. Generate a white-noise source z_j and apply the coloring filter F_{c_j} followed by a Poisson-pulse filter G_{a_j}:

$$\epsilon_j = G_{a_j}(F_{c_j}(z_j)). \tag{2}$$

To construct the filter G_{a_j}, first, a sequence of spikes is generated by a homogeneous Poisson process. Here, we denote the time indexes of the spikes as $s^{(1)}, s^{(2)}, \cdots, s^{(K)}$, as shown in Figure 2 (C1). Second, at each spike $s^{(k)}$, duration $d^{(k)}$ is sampled from a Gaussian distribution $N(\mu, \sigma)$, where $\mu = \lambda a_j$, λ is a scaling parameter, and $\sigma = 1$. This makes a random pulse signal, as shown in Figure 2 (C2). Finally, the filter G_{a_j} is defined by

$$\left[G_{a_j}(x)\right]_t = \begin{cases} 10^{\frac{P_{min}}{20}\frac{t-s^{(k)}}{d^{(k)}}} \cdot x_t & \text{if } \exists k \ \ t \in R_k \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $[\cdot]_t$ denotes the value at time t; R_k is the range at the k-th spike, that is, $R_k = \{t : s^{(k)} \leq t < \min(s^{(k+1)}, s^{(k)} + d^{(k)})\}$; and P_{min} is a constant set to -120. This definition attenuates the noise level linearly in decibels at each spike. The definition is visualized in Figure 2 (C3).

Figure 2B shows the differences between categories. As can be seen, the noise structure changes from pulse-like noise to stable noise as the attenuation category label a increases. Therefore, pre-training on this dataset predisposes networks to focus more on time-domain characteristics.

3) Extended Poisson Colored Noise Dataset: The third dataset, namely the extended Poisson colored noise dataset, explores a different type of source signal to extend instance variability. It introduces a source type selection procedure to the second step of noise generation, where the source z_j is selected from white noise or a random pulse. The random pulse is a signal having spikes with a value of 1 on an all-zero background signal. The interval between adjacent spikes is randomly sampled from 4 to 12. The number of categories of this dataset is 100.

C. Pre-Training

After generating a noise dataset $\mathcal{E}$, it is used to pre-train a network $\mathcal{N}_\theta$. Any type of objective function for solving classification problems, such as softmax loss or Kullback–Leibler divergence loss [11], can be introduced into this step. Further, data augmentation methods, such as mixup learning [21], can also help improve the performance. For evaluation, the pre-trained network is fine-tuned on the given training dataset $\mathcal{D}_{train}$ having only a few examples of real-world environmental sounds.

IV. Experiments

This section describes the evaluation of the proposed method for few-example environmental sound recognition with three datasets: ESC-10, ESC-50, and UrbanSound8K. We describe the evaluation settings before moving on to the results.

A. Experimental Settings

The three datasets and their evaluation measures are as follows:

1) ESC-10 dataset. This dataset consists of 400 environmental recordings in 10 sound classes such as *sea-waves*, *crackling fire*, and *chainsaw*. We followed the evaluation protocol in [3] using fivefold cross-validation.

2) ESC-50 dataset. This dataset consists of 2,000 environmental recordings in 50 sound classes. We followed the evaluation protocol in [3] using fivefold cross-validation.

3) UrbanSound8K dataset. This dataset consists of 8,732 audio clips for 10 urban sound classes, such as *air conditioner*,

TABLE I
ACCURACY (%) OF SMALL-SAMPLE ENVIRONMENTAL SOUND RECOGNITION ON ESC-10, ESC-50, AND URBANSOUND8K DATASETS. COLOR5, PCOLOR50, AND EXPCOLOR50 DENOTE THE DATASETS PROPOSED FOR PRE-TRAINING.

	ESC-10				ESC-50				Urban8K			
# Training Samples	1	2	4	8	1	2	4	8	1	2	4	8
EnvNetV2 [2] (scratch)	37.90	54.90	67.75	75.70	2.52	17.24	23.44	63.58	10.86	18.38	28.81	34.85
- Color5	43.80	56.25	65.70	75.70	4.80	21.30	23.30	64.69	17.52	27.58	34.68	40.09
- PColor50	48.75	62.40	72.25	81.40	25.10	37.98	38.14	66.80	23.53	**33.75**	39.25	**49.14**
- ExPColor50	**50.55**	**67.05**	**76.65**	**84.95**	**27.66**	**39.49**	**39.39**	**67.16**	**26.50**	33.62	**40.04**	48.88
SE-EnvNetV2 (scratch)	42.55	56.30	68.70	76.20	20.40	38.87	52.66	69.43	11.43	15.68	20.64	27.74
- Color5	38.75	51.25	60.65	71.35	20.20	37.73	52.02	70.50	16.56	18.43	28.67	30.96
- PColor50	**51.60**	**62.75**	71.80	80.45	**27.97**	41.03	56.08	69.47	**26.74**	35.23	41.30	50.74
- ExPColor50	51.25	61.75	**72.75**	**80.85**	27.49	**41.33**	**56.39**	**69.95**	26.04	**35.49**	**41.43**	**50.84**

dog bark, and *gunshot*. We follow the evaluation protocol in [4] using 10-fold cross-validation.

To conduct experiments using few examples for training, the number of training samples per category was varied from one to eight. Experiments were repeated 10 times, and the results are reported here as average classification accuracy. Note that in all experiments, validation splits were fixed for a fair comparison.

Two types of network architectures, EnvNetV2 [2] and SE-EnvNetV2, were implemented. Here, SE-EnvNetV2 is a new network that introduces the squeeze-and-excitation (SE) module [22] to EnvNetV2. For pre-training, three proposed datasets were compared: Colored noise dataset with 5 classes (Color5), Poisson colored noise dataset with 50 classes (PColor50), and extended Poisson colored noise dataset with 50 classes (ExPColor50). Each dataset has 32 generated audio samples per category. For training on evaluation datasets, the data augmentation method in [11] was applied. The default hyperparameters were used in all experiments.

B. Experimental Results

Table I reports classification accuracy on the three evaluation datasets. We see that the proposed pre-training framework improves the performance for all datasets and conditions. This confirms the effectiveness of pre-training on colored noise and its extensions.

If we compare the three proposed pre-training datasets, PColor50 significantly improves the performance from Color5, and ExPColor50 yields a further slight improvement. This shows that datasets having both frequency-domain and time-domain varieties are effective for pre-training.

To analyze the optimization process, Figure 3 compares training accuracy curves with and without introduction of a pre-trained model. We see a clear difference between the two curves, indicating that optimization without pre-training (from "scratch") falls into a premature local optimal solution. This result supports our assumption that pre-training helps to avoid such solutions, and confirms the effectiveness for for few-example environmental sound recognition.

Finally, Table II confirms that our framework does not degrade the performance even if many samples are used for training. It also shows that our experiments were conducted with a high-performance baseline compared with other methods. Combining our method with other types of learning methods, such as unsupervised filter-bank learning [6], would be a promising next step to further improve the performance.

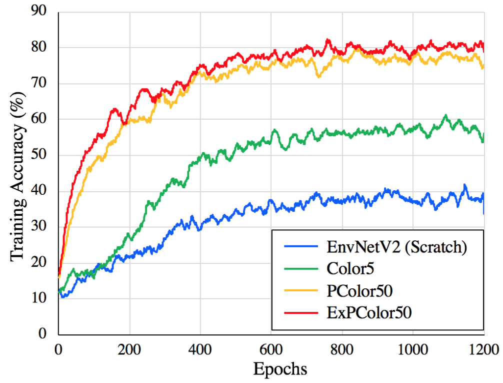

Fig. 3. Training accuracy curves on UrbanSound8K dataset using eight samples per class for training.

TABLE II
COMPARISON WITH OTHER METHODS. THE RESULTS ARE OBTAINED USING ALL SAMPLES FOR CONFIRMING THAT OUR PRE-TRAINING DOES NOT AFFECT THE PERFORMANCE EVEN WHEN MANY SAMPLES ARE USED FOR TRAINING.

Method	Input Type	Acc.
Piczak FBEs CNN [23]	Mel spectrogram	64.50
SoundNet [1]	Raw audio	74.20
ConvRBM [6]	Raw audio	78.45
Human [3]	-	81.30
TEO-GS CNN [24]	Gammatone spectrogram	81.95
PEFBEs CNN [25]	Mel spectrogram	84.15
Multi-Stream Net [9]	STFT coefficients	84.00
ConvRBM+FBEs [6]	Mel spectrogram	86.50
EnvNetV2 [2]	Raw audio	84.90
- ExPColor50	Raw audio	84.90
SE-EnvNetV2	Raw audio	86.05
- ExPColor50	Raw audio	86.35

V. CONCLUSION

This paper presents closed-form pre-training, which pre-estimates network parameters without any prior recording or annotation. We proposed three types of datasets for automatically generating and using colored noise and its extensions. We performed experiments to demonstrate the effectiveness of the proposed framework for environmental sound recognition on three publicly available datasets after fine tuning with only a few examples. A useful line of inquiry for future work would be to focus on introducing other types of noise, as well as combining the method with other types of networks.

VI. Appendix: Definition of coloring filters

Let $z = (z_1, z_2, \cdots, z_T)$ be a white-noise source. The coloring filter F_c is defined autoregressively [26] as follows:

$$[F_c(z)]_t = -\sum_{k=1}^{63} a_{c,k} \, [F_c(z)]_{t-k} + z_t, \tag{4}$$

where $a_{c,k}$ is a filter coefficient inductively defined by $a_{c,0} = 1$ and $a_{c,k} = \left(k - 1 - \frac{\beta_c}{2}\right) \frac{a_{k-1}}{k}$. The parameter β_c is defined for each color as $\beta_1 = 0$ (white), $\beta_2 = 1$ (pink), $\beta_3 = -1$ (blue), $\beta_4 = -2$ (violet), and $\beta_5 = 2$ (brown). Note that this is the standard implementation in Matlab for generating colored noise.

VII. Acknowledgement

This work was supported by the JST ACT-X Grant No. JPMJAX1905, and the JSPS KAKENHI Grant 19K22865.

References

[1] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 892–900, 2016.

[2] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2721–2725, 2017.

[3] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM International Conference on Multimedia (ACMMM)*, pp. 1015–1018, 2015.

[4] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM International Conference on Multimedia*, pp. 1041–1044, 2014.

[5] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 421–425, 2017.

[6] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification," in *Proc. Interspeech*, pp. 3107–3111, 2017.

[7] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, "Audio scene classification with deep recurrent neural networks," in *Proc. Interspeech*, pp. 3043–3047, 2017.

[8] M. Okawa, T. Saito, N. Sawada, and H. Nishizaki, "Audio classification of bit-representation waveform," in *Proc. Interspeech*, pp. 2553–2557, 2019.

[9] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," in *Proc. Interspeech*, pp. 3604–3608, 2019.

[10] Y.-H. Shen, K.-X. He, and W.-Q. Zhang, "Learning how to listen: A temporal-frequential attention model for sound event detection," in *Proc. Interspeech*, pp. 2563–2567, 2019.

[11] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[12] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2015.

[13] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[14] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 1–5, 2019.

[15] S. Gharib, K. Drossos, E. Çakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 138–142, 2018.

[16] J. Pons, J. Serrà, and X. Serra, "Training neural audio classifiers with few data," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 16–20, 2019.

[17] S. Mun and S. Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 845–849, 2019.

[18] Y. Chen and H. Jin, "Rare sound event detection using deep learning and data augmentation," in *Proc. Interspeech*, pp. 619–623, 2019.

[19] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 621–630, 2019.

[20] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end ASR with language model fusion," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6096–6100, 2019.

[21] Y. Zhu, T. Ko, and B. Mak, "Mixup learning strategies for text-independent speaker verification," in *Proc. Interspeech*, pp. 4345–4349, 2019.

[22] W. Xia and K. Koishida, "Sound event detection in multichannel audio using convolutional time-frequency-channel squeeze and excitation," in *Proc. Interspeech*, pp. 3629–3633, 2019.

[23] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2015.

[24] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel TEO-based gammatone features for environmental sound classification," *IEEE European Signal Processing Conference (EUSIPCO)*, pp. 1809–1813, 2017.

[25] R. N. Tak, D. M. Agrawal, and H. A. Patil, "Novel phase encoded mel filterbank energies for environmental sound classification," *Pattern Recognition and Machine Intelligence*, pp. 317–325, 2017.

[26] N. J. Kasdin, "Discrete simulation of colored noise and stochastic processes and $1/f^{\alpha}$ power law noise generation," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 802–827, 1995.