DESCRIPTION AND DISCUSSION ON DCASE 2024 CHALLENGE TASK 2: FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING

Tomoya Nishida[1], Noboru Harada[2], Daisuke Niizumi[2], Davide Albertini[3], Roberto Sannino[3], Simone Pradolini[3], Filippo Augusti[3], Keisuke Imoto[4], Kota Dohi[1], Harsh Purohit[1], Takashi Endo[1], and Yohei Kawaguchi[1]

[1] Hitachi, Ltd., Japan, tomoya.nishida.ax@hitachi.com
[2] NTT Corporation, Japan, noboru.harada.pv@hco.ntt.co.jp
[3] STMicroelectronics, Switzerland,
[4] Doshisha University, Japan, keisuke.imoto@ieee.org

ABSTRACT

We present the task description of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 Challenge Task 2: "First-shot unsupervised anomalous sound detection (ASD) for machine condition monitoring". Continuing from last year's DCASE 2023 Challenge Task 2, we organize the task as a first-shot problem under domain generalization required settings. The main goal of the first-shot problem is to enable rapid deployment of ASD systems for new kinds of machines without the need for machine-specific hyperparameter tunings. For the DCASE 2024 Challenge Task 2, sounds of new machine types were collected and provided as the evaluation dataset. In addition, attribute information such as the machine operation conditions were concealed for several machine types to simulate situations where such information are unavailable. We received 96 submissions from 27 teams, and an analysis of these submissions has been made in this paper. Several novel approaches, such as new ways of utilizing pre-trained models and pseudo-label classification approaches, have been used to beat the baseline system.

Index Terms— anomaly detection, acoustic condition monitoring, domain shift, first-shot problem, DCASE Challenge

1. INTRODUCTION

Anomalous sound detection (ASD) [1–7] is the task of identifying whether the sound emitted from a target machine is normal or anomalous. This leads to automatic detection of mechanical failures, which is vital in the fourth industrial revolution with AI-based factory automation. Using machine sounds for prompt detection of machine anomalies is useful for machine condition monitoring.

A major challenge concerning the application of ASD systems is that both the number and variety of anomalous samples can be inadequate in training. In 2020, we held the first ASD task in Detection and Classification of Acoustic Scenes and Event (DCASE) Challenge 2020 Task 2 [8]; "unsupervised ASD" which aimed to detect unknown anomalous sounds using only normal sound samples as training data. Following this task, handling of domain shifts was additionally tackled in the DCASE Challenge 2021 Task 2 [9] and 2022 Task 2 [10] for the wide spread application of ASD systems. Domain shifts are differences between the data in the source and target domains, which are caused by shifts in the operational conditions of the machine or environmental noise. The DCASE Challenge 2021 Task 2 [9] mainly focused on the use of domain adaptation techniques, whereas the DCASE Challenge 2022 Task 2 [10] focused on the use of domain generalization techniques.

In the DCASE Challenge 2023 Task 2 [11], "first-shot unsupervised ASD," real-world scenarios were explored even further as a "first-shot" ASD task. This is a task that requires solving UASD against completely novel machine types, without access to data from similar machine types that can be used for model training or hyperparameter tuning. This scenario is typically encountered in real-world situations where the rapid deployment of ASD systems is required and collecting a variety of training or test data is infeasible. To realize this problem setting, the evaluation dataset was created by completely new machine types unseen in the development dataset. This setup prevented participants from performing handcrafted tunings which are difficult to implement in many real-world applications. For example, hyperparameter tuning for each machine type using the development dataset or training ASD systems with the same machine type sounds has become infeasible.

To further deepen the techniques that are useful for this problem setting grounded on real-world scenarios, we designed the DCASE Challenge 2024 Task 2 "First-shot unsupervised anomalous sound detection for machine condition monitoring" by closely aligning to the problem setting established in the previous year. The main modifications from DCASE 2023 Task 2 are that the evaluation dataset is updated with new machine types unseen in the previous DCASE ASD challenges, and that attribute information such as the machine operation conditions are concealed for several machine types. The second modification concerns situations where such information is unavailable, with the aim of expanding the range of applicable scenarios in real-world settings.

2. FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION UNDER DOMAIN SHIFTED CONDITIONS

Let the L-sample time-domain observation $\boldsymbol{x} \in \mathbb{R}^L$ be an audio clip that includes sounds emitted from a machine. The goal of the ASD task is to determine a given machine as either normal or anomalous by computing an anomaly score $\mathcal{A}_\theta(\boldsymbol{x})$ using an anomaly score calculator $\mathcal{A}$ with parameters θ. The input of $\mathcal{A}$ can be the audio clip $\boldsymbol{x}$ or $\boldsymbol{x}$ with additional information such as labels indicating the operation condition of the machine. The machine is then determined

to be anomalous when $\mathcal{A}_\theta(\boldsymbol{x})$ exceeds a pre-defined threshold ϕ as

$$\text{Decision} = \begin{cases} \text{Anomaly} & (\mathcal{A}_\theta(\boldsymbol{x}) > \phi) \\ \text{Normal} & (\text{otherwise}). \end{cases} \quad (1)$$

The primary difficulty in this task is to train the anomaly score calculator with only normal sounds (unsupervised ASD). The DCASE 2020 Challenge Task 2 [8] was designed to address this issue, and all the following tasks stand on this unsupervised ASD setting.

The domain-shift problem also needs to be solved for practical applications of ASD. Domain shifts are variations in conditions between training and testing phases that change the distribution of the observed sound data. These shifts can arise from differences in operating speed, machine load, heating temperature, environmental noise, microphone arrangement, and other factors. Two domains, the **source domain** and the **target domain**, are defined: the former refers to the original condition with sufficient training data and the latter refers to another condition with only a few samples. This year's task follows the 2022 and 2023 Task 2 [10, 11] setting, where the domain information is assumed to be unknown in the test phase and anomalies from both domains have to be detected with a single threshold. In this case, domain generalization is required to achieve good performance.

To further pursue the rapid development of ASD systems in real-world scenarios, solving ASD (a) against completely novel machine types (b) with only one section of training data (c) without handcrafted tunings that depend on test data, are highly important. This is because in real-world scenarios, customers may only possess a single novel machine, and collecting test data for handcrafted tuning may be infeasible. This problem setting was named as the "first-shot problem", and the 2023 Task 2 [11] was organized based on this problem setting. Specifically, the first-shot problem was realized by adding two features to the dataset: (i) Completely different sets of machine types between the development and evaluation dataset and (ii) Only one section for each machine type. Note that until 2022 Task 2, the data provided included multiple sections for each machine type, with the development and evaluation datasets sharing the same machine types.

While solving the first-shot problem under the domain generalization setting should be sufficient for many real-world applications, the results from the previous year suggested that there is still potential for further improvement in the solutions [11]. For this reason, we designed the DCASE Challenge 2024 Task 2, "First-shot unsupervised anomalous sound detection for machine condition monitoring" by closely aligning to the problem setting designed in the previous year. The main modifications from 2023 Task 2 are that the evaluation dataset consists of newly recorded sounds of new machine types and that attribute information are concealed for several machine types. By mostly following the same problem setting as in DCASE 2023 Task 2, the organizers aim to further deepen the techniques that are useful for first-shot ASD.

3. TASK SETUP

3.1. Dataset

The data for this task comprises three datasets: **development dataset**, **additional training dataset**, and **evaluation dataset**. The development dataset includes seven machine types, whereas the additional and evaluation dataset includes nine machine types, each having one section per machine type. **Machine type** means the type of machine such as fan, gearbox, etc. **Section** is a subset or whole data within each machine type.

Each recording is a single-channel audio with a duration of 6 to 10 s and a sampling rate of 16 kHz. We mixed machine sounds recorded at laboratories with environmental noise recorded at factories and in the suburbs to create each sample in the dataset. For the details of the recording procedure, please refer to the papers on ToyADMOS2 [12] and MIMII DG [13].

The **development dataset** consists of seven machine types (fan, gearbox, bearing, slide rail, valve, ToyCar, ToyTrain), and each machine type has one section that contains a complete set of the training and test data. Each section provides (i) 990 normal clips from a source domain for training, (ii) 10 normal clips from a target domain for training, and (iii) 100 normal clips and 100 anomalous clips from both domains for the test. We provided domain information (source/target) in the test data for the convenience of participants. For four machine types (fan, bearing, valve, ToyCar), attributes that represent operational or environmental conditions are also provided in the file names and attribute csvs. For the other three machine types, attributes are concealed. The **additional training dataset** provides novel nine machine types (3D-printer, air compressor, brushless motor, hairdryer, hovering drone, robotic arm, scanner, toothbrush, ToyCircuit). Each section consists of (i) 990 normal clips in a source domain for training and (ii) 10 normal clips in a target domain for training. For five machine types (3D-printer, hairdryer, robotic arm, scanner, ToyCircuit), attributes are provided in this dataset. For the other four machine types, attributes are concealed. The **evaluation dataset** provides the test clips that correspond to the additional training dataset, e.g. data of the same machine types as the additional training dataset. Each section consists of 200 test clips, none of which have a condition label (i.e., normal or anomaly), domain information, or attribute information.

Participants must train a model for a new machine type using only one section per machine type, without hyperparameter tuning using test datasets obtained from the same machine type, and for some of the machine types, without utilizing attribute information.

3.2. Evaluation metrics

We used the area under the receiver operating characteristic curve (AUC) to evaluate overall detection performance and the partial AUC (pAUC) to measure performance in a low false-positive rate range $[0, p]$, where we set $p = 0.1$. To evaluate each system under the domain generalization setting, we compute the AUC for each domain and pAUC for each section as

$$\text{AUC}_{m,n,d} = \frac{1}{N_d^- N_n^+} \sum_{i=1}^{N_d^-} \sum_{j=1}^{N_n^+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (2)$$

$$\text{pAUC}_{m,n} = \frac{1}{\lfloor pN_n^- \rfloor N_n^+} \sum_{i=1}^{\lfloor pN_n^- \rfloor N_n^+} \sum_{j=1}^{N_n^+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (3)$$

where m and n represent the index of a machine type and a section respectively, $d \in \{\text{source}, \text{target}\}$ represents a domain, $\lfloor \cdot \rfloor$ is the flooring function, and $\mathcal{H}(y)$ returns 1 when $y > 0$ and 0 otherwise. Here, $\{x_i^-\}_{i=1}^{N_d^-}$ are the normal test clips in domain d in section n of machine type m and $\{x_j^+\}_{j=1}^{N_n^+}$ are all the anomalous test clips in section n of machine type m. N_d^-, N_n^-, N_n^+ represent the number of normal test clips in domain d, normal test clips in section n, and anomalous test clips in section n, respectively.

The official score Ω is given by the harmonic mean of the AUC and pAUC scores overall machine types and sections:

$$\Omega = h\left\{\mathrm{AUC}_{m,n,d},\ \mathrm{pAUC}_{m,n} \mid m \in \mathcal{M},\ n \in \mathcal{S}(m),\ d \in \{\mathrm{source}, \mathrm{target}\}\right\}, \quad (4)$$

where $h\{\cdot\}$ represents the harmonic mean, $\mathcal{M}$ is the set of given machine types, and $\mathcal{S}(m)$ represents the set of sections for machine type m. Specifically, $\mathcal{S}(m) = \{00\}$ for the dataset in 2024.

3.3. Baseline systems and results

The task organizers provide a baseline system based on Autoencoders (AEs), featuring two distinct operating modes. This baseline system is the same system employed as the baseline in 2023 Task 2. Although both modes employ Autoencoder for training, they diverge in the computation of anomaly scores. In this paper, we introduce the baseline system along with its detection performance. For further information, please refer to [14].

3.3.1. Autoencoder training

The AE is first trained for both operating modes. First, the log-mel-spectrograms of each training sound clips $X = [X_1, \ldots, X_T]$ are calculated, where $X_t \in \mathbb{R}^F$ for $t = 1, \ldots, T$ are the frame-wise feature vectors at frame t, $F = 128$ is the number of mel-filters and T is the number of time-frames. For the input of the AE, $P = 5$ consecutive frames taken from X are concatenated as $\psi_t = [X_t^\mathsf{T}, \ldots, X_{t+P-1}^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^D$ for each t, where $D = P \times F = 640$. The model parameters are optimized by minimizing the mean squared error (MSE) between the input ψ_t and the reconstructed output $r_\theta(\psi_t)$ for all inputs created from the training data.

3.3.2. Simple Autoencoder mode

In this mode, the anomaly score is calculated as the average of the MSE for all input features created from that sound clip, e.g.,

$$A_\theta(X) = \frac{1}{DK}\sum_{k=1}^{K} \|\psi_k - r_\theta(\psi_k)\|_2^2, \quad (5)$$

where $K = T - P + 1$, and $\|\cdot\|_2$ represents ℓ_2 norm.

3.3.3. Selective Mahalanobis mode

In this mode, the Mahalanobis distance between the system input and reconstructed feature is used to calculate the anomaly score. The anomaly score is given as

$$A_\theta(X) = \frac{1}{DK}\sum_{k=1}^{K} \min\{D_s(\psi_k, r_\theta(\psi_k)), D_t(\psi_k, r_\theta(\psi_k))\}, \quad (6)$$

$$D_s(\cdot) = \mathrm{Mahalanobis}(\psi_k, r_\theta(\psi_k), \Sigma_s^{-1}), \quad (7)$$

$$D_t(\cdot) = \mathrm{Mahalanobis}(\psi_k, r_\theta(\psi_k), \Sigma_t^{-1}), \quad (8)$$

where Σ_s^{-1} and Σ_t^{-1} are the covariance matrices of $r_\theta(\psi_k) - \psi_k$ for the source and target domain data of each machine type, respectively.

3.3.4. Results

Tables 1 show the AUC and pAUC scores for the two baselines on the development dataset. The average and standard deviations of the scores from five independent trials of training and testing are shown in the tables.

Table 1: Baseline results for development dataset.

Machine type	Mode	AUC [%]		pAUC [%]
		Source	Target	
ToyCar	MSE	66.98 ± 0.89	33.75 ± 0.81	48.77 ± 0.13
	MAHALA	63.01 ± 2.12	37.35 ± 0.83	51.04 ± 0.16
ToyTrain	MSE	76.63 ± 0.22	46.92 ± 0.80	47.95 ± 0.09
	MAHALA	61.99 ± 1.79	39.99 ± 1.37	48.21 ± 0.05
bearing	MSE	62.01 ± 0.64	61.40 ± 0.26	57.58 ± 0.32
	MAHALA	54.43 ± 0.27	51.58 ± 1.73	58.82 ± 0.13
fan	MSE	67.71 ± 0.70	55.24 ± 0.91	57.53 ± 0.19
	MAHALA	79.37 ± 0.44	42.70 ± 0.26	53.44 ± 1.03
gearbox	MSE	70.40 ± 0.58	69.34 ± 0.82	55.65 ± 0.44
	MAHALA	81.82 ± 0.33	74.35 ± 1.21	55.74 ± 0.35
slider	MSE	66.51 ± 1.66	56.01 ± 0.29	51.77 ± 0.35
	MAHALA	75.35 ± 3.02	68.11 ± 0.63	49.05 ± 1.00
valve	MSE	51.07 ± 0.88	46.25 ± 1.30	52.42 ± 0.50
	MAHALA	55.69 ± 1.44	53.61 ± 0.19	51.26 ± 0.47

4. CHALLENGE RESULTS

We received 96 submissions from 27 teams. Ten teams outperformed the simple Autoencoder baseline, and eleven outperformed the selective Mahalanobis baseline, which indicates the difficulty of the task. The number of teams outperforming the baselines was also close to that in 2023's task. Figure 1 shows the AUC values for the top 10 teams. In the source domain, many teams successfully improved the AUC values for half of the machine types, but showed lower AUC values than the baseline in the other half. As a result, the harmonic mean of the AUC values in the source domain was very close to the baselines for all teams. In the target domain, most of the top ten teams outperformed the baselines in most machine types. The order of the harmonic mean in the target domain was mostly aligned with the order of the official ranks, which means the performance on this domain was the key to achieve higher ranks.

Figure 2 compares the AUC values of the top 20 teams between the development and evaluation datasets. As can be seen, achieving high AUC values in the development dataset does not necessarily imply high AUC values in the evaluation dataset. This trend is seen especially in the first shot problem setting; The correlation coefficients between the mean AUCs of the development and evaluation dataset were higher for non-first shot tasks, e.g., 0.82 for 2021 and 0.83 for 2022, and lower for the first shot tasks, e.g., 0.62 for 2023 and 0.14 for 2024. This clarifies the difficulty of the first shot problem setting. As a result, teams that achieved high AUC values in the evaluation dataset (especially in the target domain, as noted above,) achieved higher ranks. Finally, in Figure 3, we compare the AUC values of the top 20 teams between machine types in which attribute information was provided and those in which attribute information was concealed. The number of teams that beat the baseline only for attribute-concealed machines (1) was fewer than that for attribute-available machines (6), which reveals that hiding the attribute has made the problem more challenging to some extent. Nevertheless, many high-ranking teams were able to surpass the baselines for both groups, indicating that those teams' solutions were capable of handling this new problem setting.

We summarize approaches used by top-ranked teams below.

a. Use of appropriate pre-trained models with fine-tunings

Using classification tasks such as machine type, domain or attribute classification as an auxiliary task to train a feature extractor remained to be a popular solution this year [15–18], following last year's trend [11]. Among them, several new attempts at using pre-trained models have achieved comparatively high scores this year. The 1st [15] and 2nd ranked team [16] both fine-tuned pre-trained models BEATs [19] and EAT [20] using low-rank adaptation (LoRA) [21], which may have prevented the model from overfitting by reducing the number of parameters to train. In addition, instead

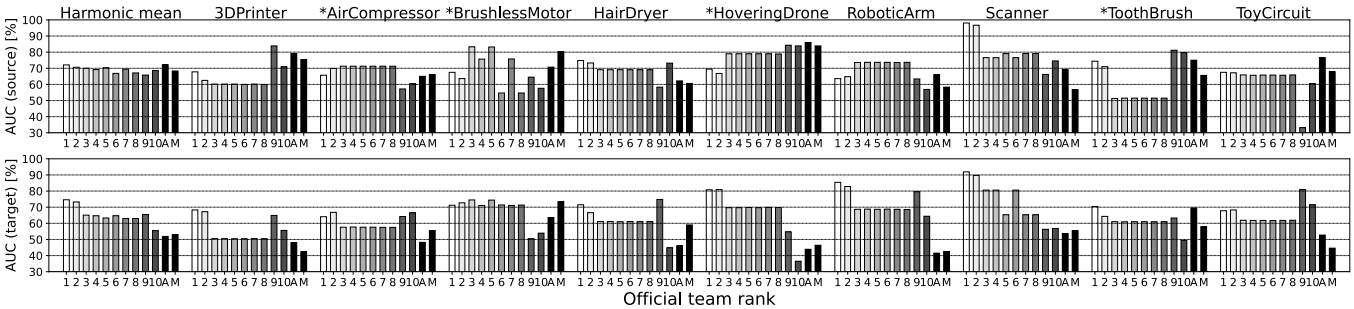

Figure 1: Evaluation results of top 10 teams in ranking. Average source-domain AUC (top) and target-domain AUC (bottom) for each machine type. Labels "A" and "M" on x-axis denote simple Autoencoder mode and selective Mahalanobis mode, respectively. "*" on machine type names indicates that attributes are hidden.

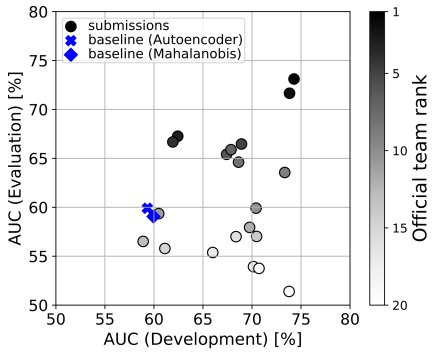

Figure 2: Comparison of average AUC for development and evaluation dataset across teams.

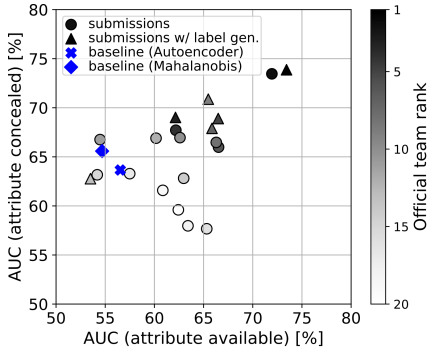

Figure 3: Comparison of average AUC for attribute available and attribute concealed machine types (development and evaluation dataset) across teams. "label gen." refers to pseudo label or auxiliary label generation approaches.

of just ensembling fine-tuned models by adding anomaly scores, both teams created a single model that has two pre-trained models in two branches and fine-tuned them simultaniously. This can automatically balance the influence of the two models on the output. Overall, enhancing ASD performance in this approach could be achieved by investigating deeper into the selection of pre-trained models and the training methodology.

b. Pseudo or auxiliary attribute labeling

Among the classification approaches that proved useful in ASD [11], attribute classification could not be used for nearly half of the machine types this year because attribute information was concealed. To deal with this situation, several teams in the top rankings generated pseudo or auxiliary labels and used them for the auxiliary classification task [15, 22–26]. The 1st ranked team [15] applied agglomerative hierarchical clustering to the audio embeddings, whereas the 9th team [25] proposed and used a bottom-up clustering method to obtain pseudo labels. The 3rd, 5th, and 13th teams [22, 23, 26] also applied clustering methods to spectrograms or statistical features. The 7th team [24] combined the training data with data from the attribute-available machines and used those attribute labels for the classification target.

As shown in Figure 3, teams using these strategies generally had higher AUC values for machines with concealed attributes, indicating the effectiveness of such strategies. However, these strategies mostly worked better only for certain machine types such as Slider or BrushlessMotor, which caused these high average AUC values. This limitation in the effective machine types might be because of the difficulty in distinguishing the sound of the target machine from the background noise only from the audio data. This difficulty can lead to wrongly created pseudo labels based on background noise differences, which does not help models learn the unique features of the target machine. To make this approach more widely effective for various machines, further investigation on how to generate labels, the usable conditions, and what assumptions can be helpful for these strategies is needed.

c. Other novel approaches

Using multiple types of input features such as the log-mel spectrogram and the power spectrum or other features has been introduced by the 4th, 7th, 9th, and several other teams [18, 24, 25, 27]. Several teams carefully selected external datasets and used them for pre-training their model. For example, the 5th and 6th team [17, 23] selected machine-related data or excluded human speech data from AudioSet [28] to make the pre-training data close to the target datasets. The 9th team applied a core-set selection method to AudioSet that selects samples with low anomaly scores as pre-training data, after manually selecting some machine-related classes [25].

5. CONCLUSION

We presented an overview of the task and analysis of the solutions submitted to the DCASE 2024 Challenge Task 2. The task's aim was to develop ASD systems that work for a novel machine type with a single section for each machine type, where the attribute information was concealed for several machine types. We discussed several new approaches that helped improve ASD performance, including ways of using pre-trained models and creating pseudo labels. We hope that all technical reports will contribute to advancements in the academic field and the industrial application of first-shot unsupervised ASD.

6. REFERENCES

[1] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," in *Proc. EUSIPCO*, 2017, pp. 698–702.

[2] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?" in *Proc. IEEE MLSP*, 2017.

[3] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma," *IEEE/ACM TASLP*, vol. 27, no. 1, pp. 212–224, Jan. 2019.

[4] Y. Kawaguchi, R. Tanabe, T. Endo, K. Ichige, and K. Hamada, "Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction," in *Proc. IEEE ICASSP*, 2019, pp. 865–869.

[5] Y. Koizumi, S. Saito, M. Yamaguchi, S. Murata, and N. Harada, "Batch uniformization for minimizing maximum anomaly score of DNN-based anomaly detection in sounds," in *Proc. IEEE WASPAA*, 2019, pp. 6–10.

[6] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE ICASSP*, 2020, pp. 271–275.

[7] H. Purohit, R. Tanabe, T. Endo, K. Suefusa, Y. Nikaido, and Y. Kawaguchi, "Deep autoencoding GMM-based unsupervised anomaly detection in acoustic signals and its hyperparameter optimization," in *Proc. DCASE Workshop*, 2020, pp. 175–179.

[8] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2020, pp. 81–85.

[9] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE Workshop*, 2021, pp. 186–190.

[10] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE Workshop*, 2022, pp. 26–30.

[11] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2023, pp. 31–35.

[12] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proc. DCASE Workshop*, 2021, pp. 1–5.

[13] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proc. DCASE Workshop*, 2022.

[14] N. Harada, N. Daisuke, T. Daiki, O. Yasunori, and Y. Masahiro, "First-shot anomaly detection for machine condition monitoring: a domain generalization baseline," in *Proc. EUSIPCO*, 2023, pp. 191–195.

[15] Z. Lv, A. Jiang, B. Han, Y. Liang, Y. Qian, X. Chen, J. Liu, and P. Fan, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[16] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Thuee system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[17] Y. Liu, "Dual-mode framework for first-shot unsupervised anomalous sound detection in machine condition monitoring," DCASE2024 Challenge, Tech. Rep., June 2024.

[18] T. Wu, J. wen, Z. Yan, and X. Cheng, "Anomalous sound detection with three-subnetworks and pre-trained models," DCASE2024 Challenge, Tech. Rep., June 2024.

[19] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2023, pp. 5178–5193.

[20] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2021.

[22] R. Zhao, K. Ren, and L. Zou, "Enhanced unsupervised anomalous sound detection using conditional autoencoder for machine condition monitoring," DCASE2024 Challenge, Tech. Rep., June 2024.

[23] L. Wang, M. Cai, J. Pan, T. Gao, and X. Fang, "Two-step anomaly detection: Integrating attribute classification and generative modeling with attribute inference for diverse machine types," DCASE2024 Challenge, Tech. Rep., June 2024.

[24] F. Chu, Y. Zhou, and M. Qian, "Unified anomaly detection for machine condition monitoring: Handling attribute-rich and attribute-free scenarios," DCASE2024 Challenge, Tech. Rep., June 2024.

[25] F. Takuya, I. Kuroyanagi, and T. Toda, "The nu systems for dcase 2024 challenge task 2," DCASE2024 Challenge, Tech. Rep., June 2024.

[26] J. Tian, H. Zhang, S. Zhang, F. Xiao, Q. Zhu, W. Wang, and J. Guan, "Self-supervised anomalous sound detection with statistical clustering and contrastive learning," DCASE2024 Challenge, Tech. Rep., June 2024.

[27] J. Yang, "Adaptive framework for first-shot unsupervised anomalous sound detection in industrial machine monitoring," DCASE2024 Challenge, Tech. Rep., June 2024.

[28] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, 2017, pp. 776–780.