ANOMALOUS SOUND DETECTION METHOD USING CONTRASTIVE LEARNING

Technical Report

Kosei Ozeki, Takeru Shiraga, Takahiko Masuzaki, Nobuaki Tanaka, and Toshiyuki Kuriyama

Mitsubishi Electric Corporation, Kamakura, Kanagawa, 2478501, Japan
Ozeki.Kosei@aj.MitsubishiElectric.co.jp
Shiraga.Takeru@ea.MitsubishiElectric.co.jp
Masuzaki.Takahiko@dc.MitsubishiElectric.co.jp
Tanaka.Nobuaki@ce.MitsubishiElectric.co.jp
Kuriyama.Toshiyuki@dx.MitsubishiElectric.co.jp

ABSTRACT

This paper presents methods for anomalous sound detection for DCASE2025 Task 2. The goal of this contest is to identify whether the sounds emitted from target machines are normal or anomaly. We implemented the following approaches: 1. Anomaly detection using a pre-trained model directly. 2. Fine-tuning the DCASE general model learned in Stage1 for individual machines. 3. Implementing the flow of approach 2 with data augmentation using additional data (clean machine data or noise-only data). 4. Performing sound source separation of operation sounds and noise, followed by implementing the flows of approaches 1 or 2. As a result, our approach achieved higher accuracy compared to the baseline method in the evaluation of the development dataset.

***Index Terms*—** fine-tuning, contrastive learning, data augmentation

1. INTRODUCTION

Anomalous sound detection (ASD) is an essential technology in machine condition monitoring. DCASE2025 Task 2 [1-3] is a data challenge focused on ASD. Participants, including the authors, aim to detect anomalous sounds using only normal data for training, considering real-world applications.

The authors' group has been working on anomaly detection in time-series data and improving ASD system performance, proposing various methods [4-8]. We participated in this challenge to verify our technical level and enhance our skills. This paper proposes the algorithms and approaches we applied to DCASE2025 Task 2.

The structure of this paper is as follows. Chapter 2 explains the task and data of DCASE2025 Task 2. Chapter 3 presents the proposed algorithms. Chapter 4 shows the evaluation results. Chapter 5 concludes the paper.

2. PROBLEM DESCRIPTION

The task of DCASE2025 Task 2 is an advanced version of DCASE2024 Task 2. It includes the following five requirements, with the fifth requirement newly introduced in DCASE2025 Task 2:

1. Train the model using only normal sounds (unsupervised learning scenario).
2. Detect anomalies regardless of domain shifts (domain generalization task).
3. Train models for entirely new machine types.
4. Train the model with or without attribute information.
5. Use additional clean machine data or noise-only data to train the model (optional).

Next, we describe the provided data. There is a development dataset for seven types of machines and an evaluation dataset for eight different types of machines. Both the development and evaluation datasets contain training data and test data. Each of the training and test datasets includes source data and target data, but the source/target information is concealed in the test data of the evaluation dataset. The training data contains only normal data. The test data includes both normal and anomalous data, but the normal/anomaly information is concealed in the test data of the evaluation dataset. Additionally, supplemental data is provided for each machine in both the development and evaluation datasets. Each machine has either clean machine data or noise-only data.

3. PROPOSED ALGORITHM

We implemented the following approaches:

1. Stage1: Anomaly detection using a pre-trained model without contrastive learning.
2. Stage2_ssl: In addition to Stage1, feature representation learning through contrastive learning using samples from each device.
3. Stage2_ssl_w_supplemental_data: In addition to Stage2 _ssl, using supplemental data for data augmentation.
4. Stage1_w_audio_separation: Performing Stage1 after conducting sound source separation.
5. Stage2_ssl_w_audio_separation: Performing Stage2_ssl after conducting sound source separation.

The flowchart without sound source separation (Stage2_ssl and Stage2_ssl_w_supplemental_data) is shown in Figure 1, and the flowchart with sound source separation (Stage2_ssl_w_audio _separation) is shown in Figure 2.

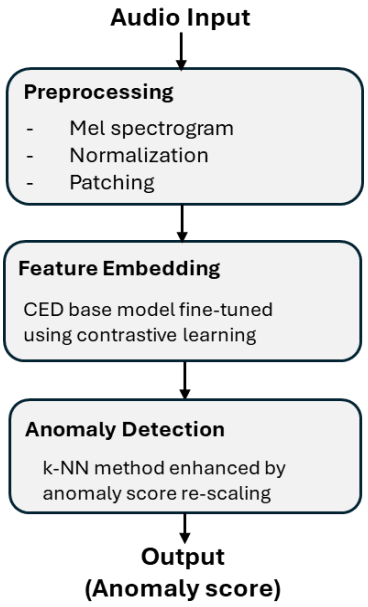

Figure 1: The flowchart without sound source separation

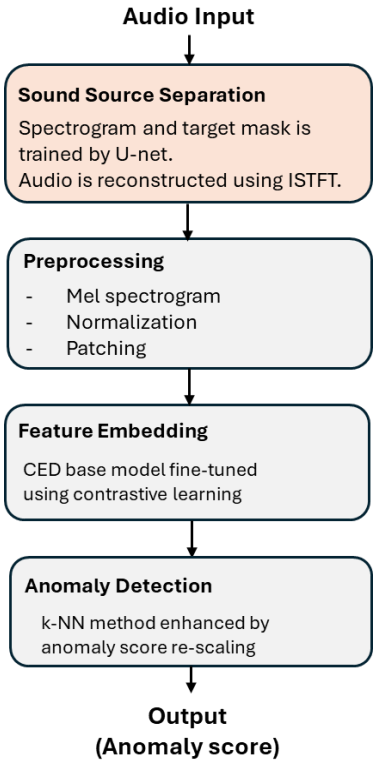

Figure 2: The flowchart with sound source separation

3.1. STAGE1

We used a pre-trained model based on CED [9] directly as a feature extractor. CED is a Vision Transformer (ViT)-based architecture proposed by Dinkel et al. for audio tagging tasks. It demonstrates strong feature extraction capabilities through pre-training on the large-scale AudioSet dataset. The pretrained model is available for download from Hugging Face.

For preprocessing, we followed the CED method and applied MelSpectrogram and patching. The feature extractor was based on CED.

For anomaly detection, we used the method by Wilkinghoff et al. [10] to handle a small number of target samples. In the following, we will refer to this method as Wilkinghoff's method. The number of samples used for k-NN rescaling, a hyperparameter, was set to 16, as suggested in the paper. We optimized the k-NN of Wilkinghoff's method using the training data. Then, we calculated the anomaly scores using the test data.

3.2. STAGE2_SSL

In Stage2_ssl, in addition to Stage1, we performed feature representation learning through contrastive learning using samples from each device (without sound source separation).

To adapt the feature space pretrained by CED to the target task, we fine-tuned the model using a SimSiam [11] based framework. Specifically, we added two network components—a projection head and a prediction head, both consisting of multilayer perceptrons (MLPs)—to the final output (embedding) of CED, and trained the entire extended model.

We created two samples from one sample with different augmentations and learned a feature space where these samples are close to each other. As a result, the trained encoder was used as the feature extractor. The data augmentation variations were as follows:

1. Gain (adjustment of gain)
2. Polarity Inversion (inverting the audio waveform vertically; it has little effect on human hearing but is used to increase data variation)
3. Pitch Shift (shifting the pitch up or down)
4. Time Stretch (changing the length of the audio without altering the pitch)
5. Add Background Noise (adding background noise from the FUSS dataset)
   - Randomly superimposing noise data provided by the FUSS dataset [12]
6. Add Background Noise (adding background noise from the additional DCASE dataset)
   - Randomly superimposing background noise data provided as supplemental data

Anomaly detection was performed in the same manner as in Stage1.

3.3. STAGE2_SSL_W_SUPPLEMENTAL_DATA

In addition to Stage2_ssl, we used supplemental data for data augmentation.

Table 1: Evaluation of the development dataset

Method		AUC_s	AUC_t	pAUC	hmean
Baseline[3]	baseline_MAHALA	**0.669**	0.509	0.532	0.556
	baseline_MSE	0.663	0.525	0.534	0.564
Proposed algorithm	stage1	0.647	0.665	0.568	0.613
	stage1_w_audio_separation	0.636	0.645	0.571	0.603
	stage2_ssl	0.661	0.675	0.575	0.624
	stage2_ssl_w_supplemental_data	0.663	**0.676**	**0.579**	**0.627**
	stage2_ssl_w_audio_separation	0.630	0.636	0.565	0.596

3.4. STAGE1_W_AUDIO_SEPARATION

In this section, we first explain sound source separation. To suppress background noise in the training and test data of the development and evaluation datasets, we designed a sound source separation model. The model used Unet [13], and we synthesized background noise with clean machine data from supplemental data to learn a hard mask for the machine data. The synthesized background noise included noise from supplemental data and audioset [14] data. Figure 3 shows the results of sound source separation for an unseen machine type (HomeCamera).

We performed Stage1 using the data processed with sound source separation as described above.

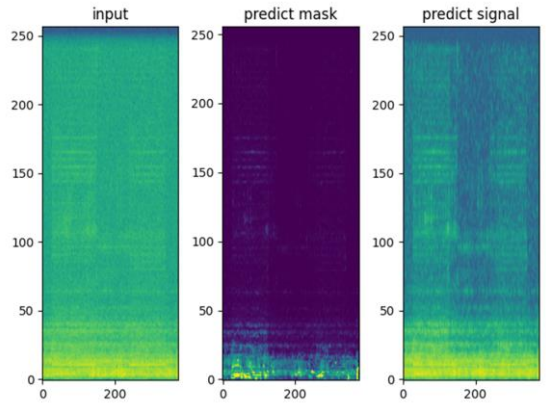

Figure 3: The results of the sound source separation process (HomeCamera).

3.1. STAGE2_SSL_W_AUDIO_SEPARATION

We performed Stage2_ssl using the data processed with sound source separation as input.

4. EVALUATION

Table 1 shows the evaluation results of the machine average in the development dataset. Here, hmean is the harmonic mean of three metrics (AUC_s, AUC_t, pAUC) and is listed as the overall score. None of the proposed methods (five in total) fell below the baseline hmean. Since the Wilkinghoff's method was introduced, even Stage1, which did not perform fine-tuning of the feature space, exceeded the baseline performance. The method with the highest overall accuracy for the development dataset was Stage2_SSL_w_supplemental_data (contrastive learning + data augmentation with additional data). Although the results of sound source separation appear lower when averaged across machines, there were cases where individual machines exceeded Stage2_SSL_w_supplemental_data, making it difficult to draw a definitive conclusion. Especially, machines provided with clean machine data tended to have higher accuracy.

The four submitted models are Stage1, Stage2_ssl, Stage2_ssl _w_supplemental_data, and Stage2_ssl_w_audio_separation.

5. CONCLUSION

In this paper, we introduced anomaly sound detection methods for DCASE2025 Task 2.

We implemented the following approaches: First, we used a pre-trained CED-based model directly as a feature extractor. To handle a small number of target samples, we employed Wilkinghoff's method (Stage1). Next, we performed feature representation learning through contrastive learning using samples from each device (Stage2_ssl). We also used supplemental data for data augmentation (Stage2_ssl_w_supplemental_data). Furthermore, we conducted sound source separation and then performed the previously mentioned feature representation learning, Stage2_ssl (Stage2_ssl_w_audio_separation).

In the evaluation of the development dataset, the highest overall accuracy was achieved when using supplemental data for data augmentation. However, for some machines, higher accuracy was obtained when sound source separation was performed.

6. REFERENCES

[1] 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 Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Barcelona, Spain, November 2021, pp. 1-5.

[2] 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 Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), Nancy, France, November 2022.

[3] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, “First-shot anomaly detection for machine condition monitoring: A domain generalization baseline,” Proceedings of 31st European Signal Processing Conference (EUSIPCO), pp. 191-195, 2023.

[4] T. Nakamura, M. Imamura, R. Mercer, & E. Keogh, “MERLIN: Parameter-Free Discovery of Arbitrary Length Anomalies in Massive Time Series Archives,” 2020 IEEE international conference on data mining (ICDM). doi:10.1109/ICDM50108.2020.00147

[5] T. Nakamura, R. Mercer, M. Imamura, & E. Keogh, “MERLIN++: parameter-free discovery of time series anomalies,” Data Mining and Knowledge Discovery, vol. 37, pp. 670-709. doi:10.1007/s10618-022-00876-7

[6] N. Tanaka, T. Shiraga, and Y. Itani, “Improving Anomalous Sound Detection by Distance Matrix-Based Visualization of Measurement Flaws,” Vol. 4 No. 1 (2023): Proceedings of the Asia Pacific Conference of the PHM Society 2023. doi:10.36001/phmap.2023.v4i1.3754

[7] T. Shiraga, H. Makimoto, et al. "Improving valvular pathologies and ventricular dysfunction diagnostic efficiency using combined auscultation and electrocardiography data: A multimodal AI approach." *Sensors* 23.24 (2023): 9834.

[8] H. Makimoto, T. Shiraga, et al. "Efficient screening for severe aortic valve stenosis using understandable artificial intelligence: a prospective diagnostic accuracy study." *European Heart Journal-Digital Health* 3.2 (2022): 141-152.

[9] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, Y. Wang, “CED: Consistent ensemble distillation for audio tagging,” arXiv: 2308.11957 (2023)

[10] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern and J. L. Roux, "Keeping the Balance: Anomaly Score Calculation for Domain Generalization," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10888402.

[11] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 15745-15753, doi: 10.1109/CVPR46437.2021.01549.

[12] Scott Wisdom, Hakan Erdogan, Dan Ellis, & John R. Hershey. (2020). Free Universal Sound Separation Dataset (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3694384

[13] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar and T. Weyde, “Singing voice separation with deep U-Net convolutional networks,” In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), doi:10.23919/EUSIPCO.2019.8902810

[14] J. F. Gemmeke, D. P. W. 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,” Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780, 2017

[15] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, “Description and Discussion on DCASE 2025 Challenge Task 2: First-shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring,” Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, 2025. doi: 10.48550/arXiv.2506.10097