AUDIOCC SYSTEM FOR DCASE 2026 TASK 2: FINE-TUNED AUDIO FOUNDATION MODELS FOR NOISE-AWARE MACHINE SOUND ANOMALY DETECTION

Technical Report

Xinhu Zheng[1], Junjie Li[2], Anbai Jiang[2], Wenrui Liang[2], Tianyu Liu[2], Shuwei Zhang[2]
Yanmin Qian[1], Xie Chen[1], Cheng Lu[3], Pingyi Fan[2], Wei-Qiang Zhang[2], Jia Liu[2]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Tsinghua University, Beijing, China
[3] North China Electric Power University, Beijing, China
Email: zhengxh24@sjtu.edu.cn

ABSTRACT

This report presents the AudioCC submission to DCASE 2026 Challenge Task 2 on noise-aware machine sound anomaly detection. Our submission studies how self-supervised audio foundation models can be adapted to machine-condition monitoring through task-oriented fine-tuning, angular-margin representation learning, and distance-based anomaly scoring. We submit four systems consisting of two single-model systems and two score-fusion systems. The submitted systems achieve a best harmonic mean of 65.33% on the development dataset.

***Index Terms*—** DCASE Challenge, anomaly detection, audio foundation model, fine-tuning, noise-aware, score fusion

1. INTRODUCTION

The DCASE 2026 Challenge Task 2 [1] addresses noise-aware machine sound anomaly detection for machine condition monitoring. The task builds on prior machine-sound ASD datasets and baselines, including ToyADMOS2 [2], MIMII DG [3], and first-shot ASD for machine condition monitoring [4]. In this setting, models must detect unknown faults using only normal training recordings, while remaining robust to source-target domain differences and channel-dependent noise.

A central focus of the 2026 task is the paired dual-channel recording setup. Each recording provides a near-field channel and a far-field channel that observe the same machine under different acoustic conditions. The far-field channel may contain stronger environmental interference, so a submitted system must account for channel-dependent noise when producing machine-wise anomaly scores.

The difficulty of this task is concentrated in three aspects:

- **Limited normal-condition supervision**. The training data contain normal recordings only. The model must therefore learn a compact description of normal operation rather than a direct boundary between normal and abnormal sounds.
- **Domain and noise variation**. Normal sounds may shift across machine instances, sections, recording conditions, and microphone positions. A useful representation must preserve machine-specific cues while reducing sensitivity to incidental acoustic variation.
- **Score comparability**. Different backbones and fine-tuning objectives produce embeddings with different score scales. Reliable submission systems require normalized anomaly scores before model combination.

Self-supervised audio models provide a natural starting point for this task because they encode broad acoustic regularities before task-specific adaptation. Recent DCASE submissions [5, 6] and ASD studies [7, 8] have shown that adapting pre-trained audio models can improve generalization under limited labeled information [9, 10]. The main question in the AudioCC submission is therefore not whether a foundation model can be used, but how the model should be fine-tuned and scored for a noise-aware ASD benchmark.

We focus on three design choices. First, the backbone is adapted with classification-oriented objectives that encourage compact representations for normal operating conditions. Second, angular-margin learning [11] is used to make the embedding space more suitable for nearest-neighbor and covariance-based anomaly scoring. Third, independently trained fine-tuned variants are combined at the score level, allowing the final system to benefit from complementary training objectives, temporal contexts, and score distributions.

The rest of this report is organized as follows. Section 2 describes the fine-tuning and scoring pipeline. Section 3 presents the four submitted systems. Section 4 reports the development-set performance.

2. METHODS

2.1. Backbone Adaptation

We employ pre-trained audio foundation models as feature extractors. The submitted systems use BEATs [12] and FISHER-small [13] backbones. Audio clips are converted into spectral features and passed through the backbone encoder. The frame-level outputs are aggregated into utterance-level embeddings by a lightweight pooling head, and the embedding head is fine-tuned together with selected backbone parameters.

The fine-tuning objective is formulated as a proxy classification task over normal operating conditions. Depending on the available metadata, the target may correspond to machine section, attribute group, or a composite label. This proxy task is not used to predict anomalies directly. Instead, it shapes the embedding space so that

Table 1: Results of four submitted systems on the development set

Machine	Metric	System 1	System 2	System 3	System 4
bearingEmu	AUC_s	63.70	60.52	64.82	64.26
	AUC_t	63.30	57.80	62.44	60.18
	pAUC	54.58	53.95	54.47	54.16
	hmean	60.22	57.30	60.24	59.24
fan	AUC_s	81.70	75.02	82.12	79.92
	AUC_t	67.90	70.38	65.34	61.98
	pAUC	62.37	61.74	61.95	62.21
	hmean	69.77	68.59	68.77	67.08
gearboxEmu	AUC_s	75.68	77.86	77.78	76.94
	AUC_t	82.18	75.38	79.50	76.64
	pAUC	64.26	58.47	69.58	69.21
	hmean	73.27	69.42	75.36	74.09
sliderEmu	AUC_s	59.56	57.42	58.32	60.50
	AUC_t	59.78	52.48	62.88	62.10
	pAUC	51.58	51.11	51.11	50.79
	hmean	56.71	53.54	57.02	57.34
ToyCar	AUC_s	67.22	78.04	68.94	67.28
	AUC_t	68.28	64.02	69.92	76.08
	pAUC	60.26	54.16	62.26	64.11
	hmean	65.05	63.97	66.86	68.80
ToyCarEmu	AUC_s	71.70	59.72	72.40	69.06
	AUC_t	81.62	75.90	84.52	83.22
	pAUC	60.58	56.05	61.32	52.21
	hmean	70.25	62.81	71.51	65.72
valveEmu	AUC_s	63.02	65.38	64.62	64.40
	AUC_t	69.86	81.48	70.36	68.84
	pAUC	51.26	61.11	52.37	51.47
	hmean	60.37	68.29	61.50	60.63
Overall	AUC_s	68.94	67.71	69.86	68.91
	AUC_t	70.42	68.21	70.71	69.86
	pAUC	57.84	56.66	59.01	57.74
	hmean	64.58	62.89	**65.33**	64.24

Results are reported for the submitted file-level scores. Overall AUC_s, AUC_t, and pAUC are arithmetic means over the listed machines, while Overall hmean is the official harmonic mean. AUC_s and AUC_t denote source and target domain AUC, respectively.

normal samples from comparable conditions form compact reference regions.

The use of foundation models is motivated by two practical constraints. First, the training set for each machine condition is small relative to the diversity of acoustic events that can appear in deployment. A pre-trained audio model supplies a general acoustic prior before machine-specific adaptation. Second, the final detector is distance-based, so the representation must remain geometrically stable after fine-tuning. We therefore keep the adaptation objective simple and use the backbone mainly to provide robust embeddings rather than to build an end-to-end anomaly classifier.

2.2. Fine-Tuning Strategies

The AudioCC systems explore several fine-tuning strategies while keeping the inference pipeline simple. The first strategy is full or partial task adaptation of a pre-trained model. This strategy updates the representation toward machine sounds while preserving the broad acoustic knowledge learned during pre-training. The second strategy is angular-margin training, where class separation is encouraged in the normalized embedding space. The third strategy is duration and augmentation variation, including spectral masking augmentation [14], which changes the temporal context seen by the model and reduces sensitivity to local noise artifacts.

Angular-margin losses are useful in this benchmark because the downstream detector relies on distances in the embedding space. If the embeddings of normal conditions are loose or heavily overlapping, distance-based anomaly scores become unstable. The margin-based objective reduces this problem by making the proxy classes more compact and more separated. The submitted variants use a common distance-based scoring interface, so their differences mainly reflect the representation and fine-tuning objective choices.

2.3. Anomaly Scoring

After fine-tuning, embeddings are extracted for the normal training recordings and the test recordings. Each machine type is scored independently. For a test sample, the anomaly score is computed by comparing its embedding to the normal reference set of the same machine type and section. We use distance-based scoring, including nearest-neighbor and covariance-aware variants, because these methods do not require anomalous training samples and can be applied consistently across machine types.

For score-fusion systems, branch scores are normalized before combination because different branches may produce scores with different numerical ranges. A single fixed score-combination configuration is used for the submitted files. Single-system submissions use the selected detector scores directly under the same evaluation interface. Binary decisions are obtained by applying the corresponding operating threshold to the anomaly scores.

2.4. Dual-Channel Scoring

The AudioCC systems handle the paired-channel setting through the data organization and scoring interface. During embedding extraction and scoring, the channel identity is kept together with the machine type and section, so normal references and test samples are compared under the same channel condition rather than mixing near-field and far-field recordings.

For the final AudioCC submissions, the submitted file-level scores use the near-field reference-channel stream. The channel information is used internally during machine-wise scoring, and the selected scores are then mapped back to the official file-level submission format. The ensemble systems combine branch scores after this scoring step, using the same scoring interface as the single systems.

2.5. Score-Level Fusion

The two ensemble systems combine normalized anomaly scores from multiple fine-tuned branches. Fusion is performed by a linear weighted sum of branch scores. The branch weights are selected with Bayesian optimization on the development set to maximize the official harmonic-mean criterion. We use score fusion rather than feature fusion because it keeps branches with different losses, durations, and detectors in a common output space.

The resulting systems intentionally cover different degrees of complexity: a compact BEATs single system, a compact FISHER-small single system, a focused BEATs fusion, and a broader BEATs fusion. This design makes it possible to compare single-backbone adaptation with multi-branch fine-tuning without changing the evaluation protocol.

Systems 1 and 2 serve as backbone-level references. They show how two foundation-model families behave under the same general scoring interface. Systems 3 and 4 then test whether multiple fine-tuned variants of the stronger representation family provide complementary score evidence. This organization compares compact single-model adaptation with increasingly broad score-level fusion.

3. SUBMITTED SYSTEMS

We submit four systems, comprising two single-model systems and two score-level fusion systems.

- **System 1**: a single BEATs backbone system with angular-margin fine-tuning and distance-based anomaly scoring.
- **System 2**: a single FISHER-small backbone system with nearest-neighbor distance scoring.
- **System 3**: a fusion system composed of three BEATs variants with Mahalanobis-style anomaly scoring.
- **System 4**: a larger fusion system composed of seven BEATs variants covering additional duration and objective choices, also using Mahalanobis-style anomaly scoring.

4. EXPERIMENT RESULTS

Detection performance is evaluated using the area under the ROC curve (AUC) for both source and target domains, partial AUC (pAUC) in the false positive rate range of 0–0.1, and their harmonic mean. Table 1 summarizes the per-machine and overall performance of all four submitted systems.

5. CONCLUSION

This report presented the AudioCC submission to DCASE 2026 Task 2. The submission emphasizes task-oriented fine-tuning of self-supervised audio foundation models for noise-aware machine sound anomaly detection. The best submitted system is a BEATs-based score fusion system and achieves a harmonic mean of 65.33% on the development set.

6. REFERENCES

[1] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2026 challenge task 2: Noise-aware unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2606.01578*, 2026.

[2] 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.

[3] 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.

[4] 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.

[5] 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," *Challenge Detect. Class. Acoust. Scenes Events (DCASE Challenge), Tech. Rep*, 2024.

[6] X. Zheng, A. Jiang, B. Han, S. Zhang, W.-Q. Zhang, X. Chen, C. Lu, P. Fan, J. Liu, and Y. Qian, "Sjtu-aithu system for dcase 2025 anomalous sound detection challenge," DCASE2025 Challenge, Tech. Rep., June 2025.

[7] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," in *Interspeech 2024*, 2024, pp. 107–111.

[8] A. Jiang, X. Zheng, B. Han, Y. Qiu, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Adaptive prototype learning for anomalous sound detection with partially known attributes," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[9] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 969–974.

[10] B. Han, A. Jiang, X. Zheng, W.-Q. Zhang, J. Liu, P. Fan, and Y. Qian, "Exploring self-supervised audio models for generalized anomalous sound detection," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[12] 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," *arXiv preprint arXiv:2212.09058*, 2022.

[13] P. Fan, A. Jiang, S. Zhang, X. Zheng, Z. Lv, B. Han, W. Liang, J. Li, W.-Q. Zhang, Y. Qian, X. Chen, and J. Liu, "Fisher: A foundation model for multimodal industrial signal comprehensive representation," *IEEE Transactions on Industrial Informatics*, pp. 1–12, 2026.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech 2019*, 2019, pp. 2613–2617.