ANOMALOUS SOUND DETECTION SYSTEM FOR DCASE 2026 TASK 2 USING DUAL-CHANNEL SPECTRAL SUBTRACTION AND EFFICIENT AUDIO TRANSFORMER

Technical Report

Fan Chu[1]*, Mengui Qian*[1]

[1] National Intelligent Voice Innovation Center, Hefei, China,
fanchu@nivic.cn, qianmengui@my.swjtu.edu.cn

ABSTRACT

This report outlines our approach to noise-aware first-shot unsupervised anomalous sound detection for machine condition monitoring, developed for DCASE 2026 Task 2. Given the constraint of only having normal operational data, alongside the complexities of variable audio durations and the availability of two-channel recordings captured at different distances, our method focuses on leveraging dual-channel signal enhancement and a pre-trained Efficient Audio Transformer (EAT) for robust anomaly detection.

Key components of our approach include applying spectral subtraction using the distant microphone as a noise reference for effective denoising, standardizing heterogeneous audio lengths to 16 seconds via audio looping with cross-fading to suppress padding artifacts, and extracting acoustic features via an EAT backbone fine-tuned with classification objectives to enhance generalization to unknown domains and complex acoustic environments. Anomalies are detected using a K-Nearest Neighbors (KNN)-based method by measuring the distance between each test sample embedding and its nearest neighbors in the training set; greater distances imply higher anomaly likelihood.

Our approach achieved notable performance on the development set, demonstrating its effectiveness. The harmonic mean of the AUC for the target domain was 69.52% and for the source domain was 70.61%. Additionally, the harmonic mean of the Partial AUC values (p=0.1) was 56.84%. These results underscore the robustness and applicability of our methodology in detecting anomalous sounds in various operational contexts.

Index Terms— Anomalous sound detection, spectral subtraction, Efficient Audio Transformer, machine condition monitoring

1. INTRODUCTION

The DCASE 2026 Challenge Task 2 focuses on "first-shot" unsupervised anomalous sound detection (UASD) under noisy and domain-shifted conditions [1]. The task simulates a rapid-deployment scenario where systems must detect anomalies for completely novel machine types using only a single section of normal training data, without the possibility of test-driven hyperparameter tuning. Furthermore, systems must handle domain shifts between source domains and target domains [2, 3].

A key update in the 2026 challenge is the focus on noise robustness through a "noise-aware" setting. Unlike previous years that provided single-channel audio, this year's dataset provides two-channel recordings simultaneously captured by microphones placed near and far from the target machine. The dataset contains 16 kHz audio clips with durations varying from 6 to 16 seconds across different machine types.

To address the noise interference and the heterogeneous audio lengths, we propose a system that integrates traditional signal processing with a deep pre-trained model. We explicitly utilize the far-field microphone to perform spectral subtraction, effectively reducing environmental noise. Subsequently, we standardize the variable-length audio clips to a fixed length using a looping and cross-fading strategy. For feature representation, we utilize the Efficient Audio Transformer (EAT) [4], fine-tuning it with classification objectives to extract robust embeddings, followed by a standard K-Nearest Neighbors (KNN) backend for anomaly scoring.

2. PROPOSED SYSTEM

2.1. Dual-Channel Spectral Subtraction

In real-world industrial deployments, target machine sounds are frequently corrupted by high-level, non-stationary factory ambient noise. The DCASE 2026 Task 2 dataset explicitly addresses this by providing two-channel recordings simultaneously captured at different distances [1]. Mechanistically, the microphone placed close to the machine (Channel 1) captures a high-energy mixture of direct machine sounds and environmental noise, whereas the distant microphone (Channel 2) captures predominantly ambient noise with significantly attenuated target machine components due to acoustic wave propagation loss.

Based on this spatial configuration, we implement an acoustic front-end utilizing a modified spectral subtraction algorithm under the additive noise assumption to isolate the intrinsic mechanical signatures[5]. Let $x_1(t)$ and $x_2(t)$ denote the discrete time-domain signals from the near and distant channels, respectively, sampled at 16 kHz. Both signals are transformed into the time-frequency domain via the Short-Time Fourier Transform (STFT) using a Hann window with a frame length (N_{FFT}) of 512 samples and a hop size of 256 samples:

$$S_1(f,\tau) = \text{STFT}\{x_1(t)\}, \quad S_2(f,\tau) = \text{STFT}\{x_2(t)\} \tag{1}$$

where f and τ represent the frequency bin index and time frame index, respectively. The magnitude spectrum of the distant channel $|S_2(f,\tau)|$ is treating as the instantaneous noise profile and subtracted directly from the primary magnitude spectrum $|S_1(f,\tau)|$. To suppress over-subtraction artifacts and avoid negative power spectral values that cause musical noise, a half-wave rectification

threshold is applied:

$$|\hat{S}(f,\tau)| = \begin{cases} |S_1(f,\tau)| - \beta|S_2(f,\tau)|, & \text{if } |S_1(f,\tau)| > \beta|S_2(f,\tau)| \\ \gamma|S_1(f,\tau)|, & \text{otherwise} \end{cases} \tag{2}$$

where $\beta = 1.0$ represents the over-subtraction factor and $\gamma = 0.1$ serves as the spectral floor parameter[6]. Finally, the enhanced time-domain signal $\hat{x}_1(t)$ is reconstructed by performing the inverse STFT (iSTFT), combining the rectified magnitude spectrum with the unmodified phase spectrum $\angle S_1(f,\tau)$ of the primary near-source channel:

$$\hat{x}_1(t) = \text{iSTFT}\left\{|\hat{S}(f,\tau)| \cdot e^{j\angle S_1(f,\tau)}\right\} \tag{3}$$

2.2. Variable-Length Audio Alignment via Cross-Faded Looping

The 2026 task presents a significant engineering hurdle due to the heterogeneous duration of recordings, which vary from 6 to 16 seconds across machine types (e.g., ToyDrone features 16-second recordings, while other types consist of shorter 6-second or 10-second clips) [1]. Batch processing within Transformer-based neural architectures strictly requires uniform tensor dimensions.

Standard zero-padding severely skews the global energy distribution and shifts the statistics of the acoustic frames, creating artificial silent regions that degrade domain generalization. Conversely, simple cyclical replication introduces phase discontinuities and sharp waveform mismatches at the splicing junctions. In the frequency domain, these boundary steps generate broadband high-frequency transient splatters. Because self-supervised deep models are highly sensitive to sudden energy mutations, the Transformer layers tend to mistake these artificial boundary transients for mechanical faults, leading to catastrophic false positives[7].

To resolve this dimensional mismatch without injecting spectral artifacts, we employ a time-domain audio looping strategy integrated with linear cross-fading. All signals shorter than 16 seconds are cyclically duplicated until they exceed the target length, and are then truncated precisely at 16.0 seconds. For each splicing junction occurring at time boundary T_{splice}, a transition window of $T_{\text{fade}} = 0.1$ seconds is established. Within this region, the tail of the preceding segment and the head of the succeeding segment are linearly blended:

$$x_{\text{aligned}}(t) = \left(1 - \frac{t}{T_{\text{fade}}}\right) x_{\text{tail}}(t) + \left(\frac{t}{T_{\text{fade}}}\right) x_{\text{head}}(t), \quad t \in [0, T_{\text{fade}}] \tag{4}$$

This operation guarantees C^0 continuity of the waveform at the loop boundaries, smoothing out artificial impulses and preserving the stationarity of the machine's true sound.

2.3. Efficient Audio Transformer (EAT) Backbone and Fine-Tuning

For high-level acoustic feature extraction, we utilize the Efficient Audio Transformer (EAT) backbone [4]. EAT is a self-supervised transformer architecture consisting of 88 million parameters, pre-trained on the massive AudioSet-2M dataset. This pre-training gives the model strong universal feature extraction capabilities and high resilience against domain shifts.

The front-end enhanced, length-normalized 16-second waveforms are converted into 128-dimensional log-Mel spectrograms. Following the standard vision-transformer processing pipeline, the spectrogram matrix is serialized into non-overlapping patches of size 16×16 in the time-frequency plane. These patches are flattened, combined with temporal and frequency positional encodings, and fed into the EAT encoder.

To bridge the domain gap between general environmental audio scenes and highly specific industrial machine sounds under the "first-shot" constraint, the pre-trained backbone must be fine-tuned. We append a linear classification layer to the average-pooled output of the EAT encoder blocks. The network is then fine-tuned using two parallel strategies depending on the metadata configurations specified in the task rules:

1. **Attribute-Available Condition**: For machine types with disclosed metadata (e.g., fan, gearbox, ToyCarEmu), the system optimizes an Additive Angular Margin Loss (ArcFace)[8] head to classify specific operational attributes (e.g., speed, load, or background subtype).
2. **Attribute-Concealed Condition**: For machines whose attributes are completely concealed (e.g., bearingEmu, sliderEmu, valveEmu), explicit target labels are unavailable. In these scenarios, we construct composite attributes by combining the specific machine type identifier with its corresponding domain label (source or target). The network is then fine-tuned using the same Additive Angular Margin Loss (ArcFace) to classify these joint condition labels. This strategy forces the model to learn representations that are discriminative of both the machine's intrinsic mechanics and its operational domain, even in the absence of fine-grained metadata.

2.4. KNN Anomaly Scoring Backend

During the evaluation phase, the classification layer used for fine-tuning is removed. Test audio clips undergo the identical pre-processing steps—dual-channel spectral subtraction and cross-faded looping—and are passed through the fine-tuned EAT encoder to extract a robust, fixed-length acoustic embedding vector v_{test}.

Unsupervised anomaly detection is executed by measuring distances in the latent feature space via the K-Nearest Neighbors (KNN) algorithm[9]. Rather than relying on strict, definitive cluster boundaries, this approach evaluates the relative spatial distribution between the normal training samples and the incoming test samples. By calculating the distance from each test embedding to its nearest neighbors in the training set, the system captures the local density of normal states; anomalous samples inherently exhibit greater spatial deviations from this established training distribution.

For each evaluation sample v_{test}, we independently compute its cosine distance to the nearest normal training neighbor in the source domain (D_{source}) and the target domain (D_{target}). To mitigate the inherent distributional discrepancies and scoring biases between the two domains, we apply domain-wise z-score normalization to these distances:

$$\hat{D}_{\text{source}} = \frac{D_{\text{source}} - \mu_{\text{source}}}{\sigma_{\text{source}}}, \quad \hat{D}_{\text{target}} = \frac{D_{\text{target}} - \mu_{\text{target}}}{\sigma_{\text{target}}} \tag{5}$$

where μ and σ represent the mean and standard deviation of the nearest-neighbor distances calculated within the respective domain's training set.

Finally, the overall anomaly score $A(v_{\text{test}})$ for the test sample is determined by taking the minimum of the two normalized distances:

$$A(v_{\text{test}}) = \min(\hat{D}_{\text{source}}, \hat{D}_{\text{target}}) \tag{6}$$

This domain-wise normalization and minimization strategy effectively prevents the target domain samples from being overwhelmingly classified as anomalies simply due to the environmental domain shift.

3. EXPERIMENTAL RESULTS

We evaluated our system using the DCASE 2026 Task 2 development dataset. Performance is measured by the Area Under the Receiver Operating Characteristic Curve (AUC) and the partial AUC (pAUC, with $p = 0.1$).

Table 1: Anomaly detection results for different machine types

Machine Type	Metric	Baseline (MSE)	Baseline (MAHALA)	Our System
ToyCarEmu	AUC(source)	**69.62%**	69.49%	67.20%
	AUC(target)	61.20%	66.62%	**82.80%**
	pAUC	55.89%	53.47%	**59.42%**
ToyCar	AUC(source)	75.62%	77.28%	**78.84%**
	AUC(target)	37.87%	53.17%	**69.74%**
	pAUC	54.03%	58.25%	**59.37%**
bearingEmu	AUC(source)	62.34%	**65.92%**	58.44%
	AUC(target)	59.56%	**62.28%**	53.08%
	pAUC	59.85%	**60.42%**	52.16%
fan	AUC(source)	61.45%	60.00%	**81.82%**
	AUC(target)	46.94%	45.09%	**66.58%**
	pAUC	53.33%	52.29%	**57.53%**
gearboxEmu	AUC(source)	68.23%	74.48%	**74.82%**
	AUC(target)	49.78%	52.74%	**76.14%**
	pAUC	52.94%	53.97%	**55.37%**
sliderEmu	AUC(source)	67.25%	66.36%	**67.44%**
	AUC(target)	45.05%	49.18%	**66.20%**
	pAUC	**50.38%**	50.36%	49.74%
valveEmu	AUC(source)	67.74%	56.60%	**71.22%**
	AUC(target)	68.78%	56.50%	**82.10%**
	pAUC	55.08%	50.20%	**67.79%**
All (hmean)	AUC(source)	67.19%	66.45%	**70.61%**
	AUC(target)	50.85%	54.24%	**69.52%**
	pAUC	54.36%	53.91%	**56.84%**

Table 1 shows that our system outperforms the official AE-based baselines (MSE and MAHALA) [1, 10] on most machine types. The combination of dual-channel spectral subtraction and EAT features demonstrates significant improvements in domain generalization. Specifically, the harmonic mean of the target domain AUC is improved from 54.24% (MAHALA) to 69.52%, indicating that the system is robust against operational condition variations and environmental noise.

4. CONCLUSION

In this report, we described our approach for DCASE 2026 Task 2, which leverages dual-channel spectral subtraction for spatial noise reduction, cross-faded looping for variable-length audio normalization, a fine-tuned Efficient Audio Transformer (EAT) for feature extraction, and a domain-normalized KNN backend. Experimental results confirmed that this pipeline provides robust anomaly detection performance, particularly under target domain shift conditions. Based on this unified framework, we submitted four specific system configurations. System 1 incorporates the complete pre-processing pipeline but employs Low-Rank Adaptation (LoRA) instead of Full Fine-Tuning (FFT) to update the EAT backbone. System 2 utilizes FFT and dual-channel denoising but replaces the cross-faded looping technique with standard zero-padding. System 3 relies solely on the raw primary channel, disabling dual-channel spectral subtraction while maintaining cross-faded looping and FFT. Finally, System 4 represents our comprehensive architecture, fully integrating dual-channel spectral subtraction, cross-faded looping, and FFT of the EAT backbone.

5. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, *et al.*, "Description and discussion on dcase 2026 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2606.10097*, 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," *arXiv preprint arXiv:2106.02369*, 2021.

[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," *arXiv preprint arXiv:2205.13879*, 2022.

[4] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 2024.

[5] Q. Liu, Y. Yu, B. S. Han, and W. Zhou, "An improved spectral subtraction method for eliminating additive noise in condition monitoring system using fiber bragg grating sensors," *Sensors*, vol. 24, no. 2, p. 443, 2024.

[6] M. Gupta, R. Singh, and S. Singh, "Analysis of optimized spectral subtraction method for single channel speech enhancement," *Wireless Personal Communications*, vol. 128, no. 3, pp. 2203–2215, 2023.

[7] H. Nam and Y.-H. Park, "Jitter: Jigsaw temporal transformer for event reconstruction for self-supervised sound event detection," *arXiv preprint arXiv:2502.20857*, 2025.

[8] 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*, 2019, pp. 4690–4699.

[9] T. Fujimura, I. Kuroyanagi, T. Hayashi, and T. Toda, "Anomalous sound detection by end-to-end training of outlier exposure and normalizing flow with domain generalization techniques," *DCASE2023 Challenge, Tech. Rep.*, 2023.

[10] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 191–195.