UNSUPERVISED ANOMALOUS SOUND DETECTION VIA EAT AND KALMAN-FILTERED NEAR- AND FAR-FIELD DUAL-CHANNEL DATA

Technical Report

Hao Wu[1]*, Zhansai Chang*[1]*, Pengyuan zhao*[1]*, Tianju Zhao*[1]*, Yutao Zhang*[1]*, Meng Lei*[1]*, Liang Zou*[1]*,*

[1] China University of Mining and Technology, Xuzhou, China
{haowu,zschang,zhaopengyuan888,tjzhao,yutaozhang,lmsiee,liangzou}@cumt.edu.cn

ABSTRACT

This report presents our approach for DCASE 2026 Task 2 on first-shot unsupervised anomalous sound detection. To address complex acoustic environments, we propose a dual-channel fine-tuning framework utilizing an Efficient Audio Transformer (EAT). Our model is trained on both frequency-domain Kalman-filtered near-field audio and original far-field audio. To handle varying durations, log-mel filterbank (fbank) features are zero-padded to ensure uniform dimensions. During feature extraction, representations are exclusively obtained from the filtered near-field channel and subjected to precise regularization: values below 0.2 are clamped to zero, while values above 0.5 are suppressed via a tanh function.Final anomaly scores are calculated using K-Nearest Neighbors (KNN).

Our approach achieved notable performance on the development set, demonstrating its effectiveness. The AUC for the target domain was 71.33% and for the source domain was 72.28%. Additionally, the Partial AUC values (p=0.1) for the target and source domain were 58.20%. These results underscore the robustness and applicability of our methodology in detecting anomalous sounds in various operational contexts.

***Index Terms*—** first-shot, anomalous sound detection, machine condition monitoring, Efficient Audio Transformer (EAT), Kalman filter, dual-channel fine-tuning, feature regularization

1. INTRODUCTION

Anomalous Sound Detection (ASD) serves as a critical task in machine condition monitoring, aiming to distinguish normal from abnormal machine sounds without prior knowledge of anomaly patterns. The DCASE Challenge Task 2 series focuses on identifying anomalous sounds across diverse machine types, emphasizing complexities in real-world industrial environments and challenges of domain shift.[1]

In complex industrial acoustic environments, such as factory floors, the sounds of target machines are often masked by non-stationary environmental background noise and spatial reverberation. Traditional time-domain filtering strategies face significant theoretical and computational limitations. First, due to broadband masking effects, time-domain filters optimized via global mean square error are easily dominated by high-energy, low-frequency background noise. [2]This often leads to the erroneous elimination of weak, high-frequency anomalous transients. Furthermore, spatial multipath reflections in real environments cause group delay dispersion; a single time-domain finite impulse response (FIR) filter struggles to accurately compensate for the nonlinear phase fluctuations of broadband signals, which can cause physical phase cancellation of the target signal. Finally, the computational complexity for state updates in time-domain filters is extremely high ($\mathcal{O}(L^2)$), making them less efficient for real-time tracking.

To overcome these limitations and effectively leverage the provided clean target device sounds and background noise, we introduce a front-end noise reduction architecture based on Frequency-Domain Adaptive Kalman Filtering (FDAKF).[3] By decomposing broadband signals into independent frequency sub-bands via Short-Time Fourier Transform (STFT), this method aligns spatial phases locally and applies differentiated state update strategies for low-frequency noise and high-frequency anomalies. Generally, rather than directly extracting features from the denoised audio, our approach is detailed below. First, the Kalman-filtered audio is utilized to fine-tune a dual-channel Efficient Audio Transformer (EAT) model. Subsequently, the fine-tuned model is employed to extract features from the near-field audio channel. These extracted features are then strictly regularized before calculating the final anomaly scores, thereby constructing a highly robust detection system.

2. PROPOSED METHOD

2.1. Frequency-Domain Adaptive Kalman Filtering (FDAKF)

Dual-channel acoustic data collection system consists of a near-field target microphone and a far-field reference microphone. Let $d(n)$ denote the near-end observed signal, which contains a mixture of target machine sounds and environmental noise, and $x(n)$ denote the far-end reference signal, which primarily consists of environmental noise. Through overlapping framing and windowing followed by STFT, the complex spectra of the near-end and far-end signals at time frame t and frequency sub-band f are denoted as $D(t, f)$ and $X(t, f)$ respectively. The acoustic transmission from the reference to the near-end observation in the frequency domain is approximated as a complex scalar product model:

$$D(t, f) = W(t, f) \cdot X(t, f) + V(t, f) \tag{1}$$

where $W(t, f)$ is the estimated transfer function of the acoustic path at frequency f, and $V(t, f)$ is a mixed term of the target source signal and system measurement noise.

To dynamically track the transfer function $W(t, f)$, we establish a first-order Markov random walk state-space model. The state equation and observation equation are defined as:

$$W(t + 1, f) = W(t, f) + w(t, f) \tag{2}$$

$$D(t, f) = W(t, f) \cdot X(t, f) + v(t, f) \tag{3}$$

where the process noise is $w(t,f) \sim \mathcal{N}(0, q(t,f))$ and the measurement noise is $v(t,f) \sim \mathcal{N}(0, r(t,f))$.

The filter alternates between prior prediction and posterior update. Assuming smooth evolution, the prior error covariance $P_{pred}(t,f)$ is determined by the previous posterior estimate and process noise:

$$P_{pred}(t,f) = P(t-1,f) + q(t-1,f) \tag{4}$$

The innovation residual $\nu(t,f)$ is calculated using the current near-end observation and the prior estimate of the reference spectrum:

$$\nu(t,f) = D(t,f) - W(t-1,f) \cdot X(t,f) \tag{5}$$

Combining the instantaneous power of the far-end reference signal, the innovation covariance $S_{cov}(t,f)$ is estimated as:

$$S_{cov}(t,f) = |X(t,f)|^2 \cdot P_{pred}(t,f) + r(t-1,f) \tag{6}$$

To minimize the mean square error, the dynamic Kalman gain $K(t,f)$ is computed as:

$$K(t,f) = \frac{P_{pred}(t,f) \cdot X^*(t,f)}{S_{cov}(t,f) + \epsilon} \tag{7}$$

where $(\cdot)^*$ denotes the complex conjugate and ϵ prevents numerical underflow. The posterior state correction and covariance update are then executed to complete the closed-loop iteration:

$$W(t,f) = W(t-1,f) + K(t,f) \cdot \nu(t,f) \tag{8}$$

$$P(t,f) = \max\left[(1 - K(t,f) \cdot X(t,f)) \cdot P_{pred}(t,f), \epsilon_{min}\right] \tag{9}$$

To adapt to the time-varying nature of environmental noise, we introduce an adaptive parameter tracking mechanism for sub-band process covariance $q(t,f)$ and measurement covariance $r(t,f)$. The process noise is smoothed using a first-order exponential function based on the energy of the state correction ΔW:

$$q(t,f) = \alpha_q \cdot q(t-1,f) + (1-\alpha_q) \cdot |K(t,f) \cdot \nu(t,f)|^2 \tag{10}$$

When the acoustic channel changes, the increase in ΔW drives $q(t,f)$ upward, allowing rapid tracking of new background noise. Conversely, the measurement noise is smoothly updated using the innovation power:

$$r(t,f) = \lambda_r \cdot r(t-1,f) + (1-\lambda_r) \cdot |\nu(t,f)|^2 \tag{11}$$

When high-frequency anomalous target events occur, the significant rise in innovation power increases $r(t,f)$, decreasing the Kalman gain. This conservative update prevents the filter from mistakenly eliminating target anomalous acoustic features as background noise. Finally, an inverse STFT with overlap-add is applied to the pure target spectrum $\nu(t,f)$ to reconstruct the time-domain waveform $e(n)$[4].

To intuitively demonstrate the efficacy of the proposed front-end denoising algorithm, visual spectrogram comparisons of various target machines (e.g., Fan, BearingEmu, BlowerDustCollector, and SewingMachine) before and after filtering are presented in Fig. 1 to Fig. 4. In the original near-end observation signals, the weak anomalous features of the target machines are severely masked by high-energy environmental background noise. Following the FDAKF processing, the background reverberation energy is effectively suppressed, allowing the high-frequency anomalous transient features to be clearly highlighted. This ensures a much purer acoustic data foundation for the subsequent feature extraction network.

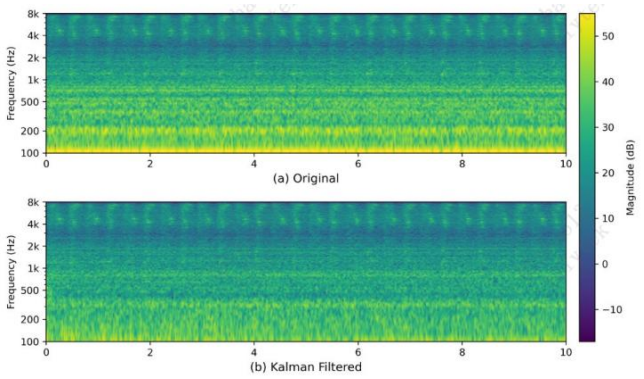

Figure 1: Spectrogram comparison of the Fan signal before and after frequency-domain adaptive Kalman filtering.

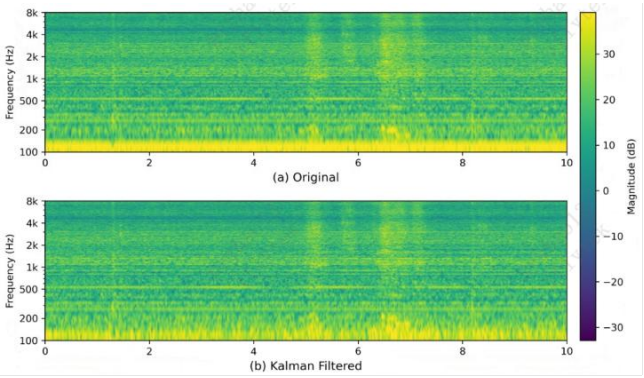

Figure 2: Spectrogram comparison of the BearingEmu signal before and after frequency-domain adaptive Kalman filtering.

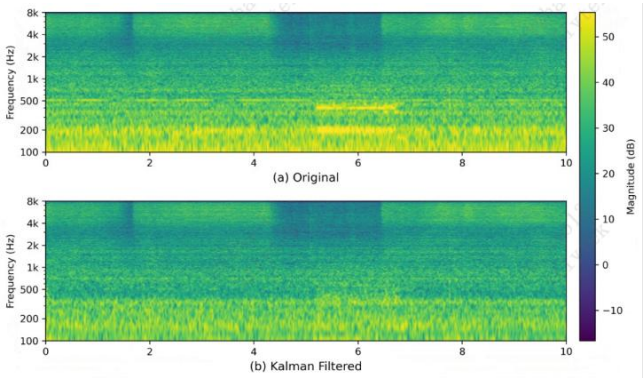

Figure 3: Spectrogram comparison of the BlowerDustCollector signal before and after frequency-domain adaptive Kalman filtering.

2.2. Preprocessing and Dual-Channel EAT Fine-Tuning

The reconstructed near-field signal $e(n)$ and the original, unfiltered far-field reference signal $x(n)$ are used jointly to fine-tune a pre-trained Efficient Audio Transformer (EAT)[5]. To ensure uniformity across varied audio clip durations, we extract log-mel filterbank (fbank) features from both channels and apply a zero-padding strategy, extending all features to represent exactly 16 seconds of

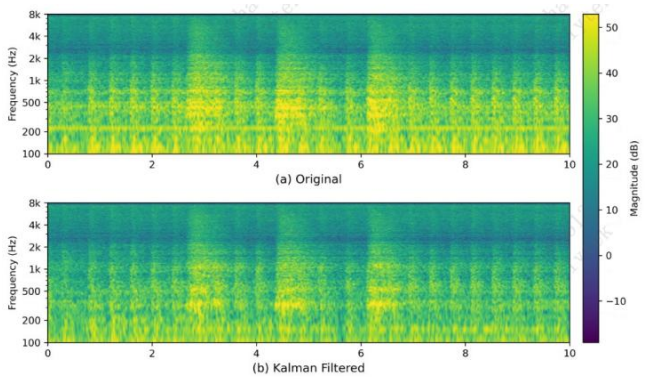

Figure 4: Spectrogram comparison of the SewingMachine signal before and after frequency-domain adaptive Kalman filtering.

audio. This dual-channel approach allows the EAT model to capture the intrinsic operational patterns of the machine while contrasting them against raw spatial acoustics.

2.3. Feature Regularization and Anomaly Scoring

During the backend inference phase, representations are exclusively extracted from the Kalman-filtered near-field channel to avoid noise contamination. Since the Efficient Audio Transformer (EAT) is pre-trained via a contrastive learning objective, its extracted embeddings inherently exhibit a high dynamic range and uncalibrated activation statistics. This often manifests as spurious low-magnitude noise and extreme activation spikes that can disproportionately dominate distance-based metrics.

To address this, we apply a strict feature regularization step: feature values less than 0.2 are clamped to zero to aggressively drop low-level artifacts, and values greater than 0.5 are suppressed using a hyperbolic tangent ($\tanh$) function to mitigate the impact of extreme outliers[6]. This pre-processing effectively regularizes the embedding distribution.

Finally, anomalies are detected using a K-Nearest Neighbors (KNN)-based method by measuring the distance between each test sample's regularized feature vector and its nearest neighbors in the training set[7]; greater distances imply higher anomaly likelihood.

3. EXPERIMENTAL RESULTS

3.1. Dataset and Evaluation Metrics

Our method was evaluated on the DCASE 2026 Task 2 development dataset[8, 9]. The performance is measured using the Area Under the Receiver Operating Characteristic curve (AUC) and the partial AUC (pAUC) calculated over a low false-positive rate range ($p = 0.1$)[10].

3.2. Performance Analysis

Visual spectrogram comparisons demonstrate the efficacy of the proposed FDAKF front-end. Across varied machine types, including Fan, BearingEmu, BlowerDustCollector, and SewingMachine, the high-energy environmental background reverberations in the original signals are heavily suppressed. Crucially, the weak, high-frequency anomalous transient features of the target machines are clearly highlighted and preserved post-filtering, providing a much purer data foundation for the downstream network.

The performance of the proposed system on the development set is summarized in Table 1. The combination of FDAKF denoising, dual-channel EAT modeling, and the strict feature regularization strategy—specifically, clamping low-level spurious noise to zero and suppressing extreme activation spikes via a tanh function—effectively stabilized the anomaly scores and significantly improved overall detection robustness.

Table 1: Experimental Results on the Development Set.

		Baseline (MSE)	Baseline (MAHALA)	Our system
ToyCarEmu	AUC(source)	69.62%	69.49%	**74.26%**
	AUC(target)	61.2%	66.62%	**85.24%**
	pAUC	55.89%	53.47%	**60.68%**
ToyCar	AUC(source)	75.62%	77.28%	**77.74%**
	AUC(target)	37.87%	53.17%	**75.68%**
	pAUC	54.03%	58.25%	**62.05%**
bearingEmu	AUC(source)	62.34%	**65.92%**	61.80%
	AUC(target)	59.56%	**62.28%**	61.98%
	pAUC	59.85%	**60.42%**	53.79%
fan	AUC(source)	61.45%	60.0%	**89.40%**
	AUC(target)	46.94%	45.09%	**68.36%**
	pAUC	53.33%	52.29%	**60.53%**
gearboxEmu	AUC(source)	68.23%	74.48%	**76.70%**
	AUC(target)	49.78%	52.74%	**82.40%**
	pAUC	52.94%	53.97%	**65.68%**
sliderEmu	AUC(source)	**67.25%**	66.36%	61.42%
	AUC(target)	45.05%	49.18%	**57.56%**
	pAUC	**50.38%**	50.36%	49.63%
valveEmu	AUC(source)	67.74%	56.6%	**72.34%**
	AUC(target)	68.78%	56.5%	**77.44%**
	pAUC	55.08%	50.2%	**58.16%**
All	AUC(source)	67.19%	66.46%	**72.28%**
	AUC(target)	50.85%	54.24%	**71.33%**
	pAUC	54.37%	53.91%	**58.20%**

4. CONCLUSION

In this technical report, we presented our submission for Task 2 of the DCASE 2026 Challenge. Our system introduces a Frequency-Domain Adaptive Kalman Filter to dynamically suppress spatial reverberation and broadband masking noise while meticulously preserving high-frequency anomaly transients. To fully leverage this front-end, Our training process is summarized below. First, the purified signal is combined with raw far-field audio in a dual-channel framework to fine-tune the Efficient Audio Transformer (EAT). Subsequently, the fine-tuned model is employed to extract representations from the near-field audio channel. By applying strict min-clamp and tanh regularizations to these extracted features prior to anomaly scoring, our method achieves highly robust detection performance. This pipeline successfully mitigates domain shift and extracts pristine machine state distributions in complex industrial acoustic environments.

5. 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," *arXiv preprint arXiv:2606.01578*, 2026.

[2] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," *arXiv preprint arXiv:2102.07820*, 2021.

[3] J. Franzen and T. Fingscheidt, "Improved measurement noise covariance estimation for n-channel feedback cancellation based on the frequency domain adaptive kalman filter," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 965–969.

[4] F. Yang, G. Enzner, and J. Yang, "Frequency-domain adaptive kalman filter with fast recovery of abrupt echo-path changes," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1778–1782, 2017.

[5] 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.

[6] K. Wilkinghoff, S. Yadav, and Z.-H. Tan, "Temporal pooling strategies for training-free anomalous sound detection with self-supervised audio embeddings," *arXiv preprint arXiv:2603.04605*, 2026.

[7] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing k-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 1, p. 113, 2024.

[8] 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.

[9] 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.

[10] 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.