MACHINE ANOMALOUS SOUND DETECTION USING NOISE-AWARE SPECTRAL SUBTRACTION AND CONFORMER BASED FEATURE LEARNING

Technical Report

Qing Zhou, Shan Li

Xi'an University of Architecture and Technology
Information and Control Engineering Dept., 13 Yanta Road, Xi'an, China
zhouq961@xauat.edu.cn, ansls@xauat.edu.cn

ABSTRACT

DCASE2026 Task 2 focuses on unsupervised anomalous sound detection for machine condition monitoring under noisy real-world environments and domain shift conditions. Strong background noise remains a major challenge, as it degrades the reliability of detection systems. To address this issue, this paper proposes a noise-aware spectral subtraction method using dual-channel recordings, followed by a Conformer-based encoder for robust feature learning. Experimental results on the task dataset demonstrate the effectiveness of the proposed approach.

Index Terms— Anomalous sound detection, conformer, spectral subtraction, self-supervised learning

1. INTRODUCTION

Anomalous sound detection (ASD) aims to automatically identify whether a target machine is normal or anomalous based on the sound it emits, playing a vital role in machine condition monitoring and predictive maintenance [5]. Recent studies have explored various strategies to improve ASD performance, including self-supervised classification, flow-based density estimation [3], contrastive learning with machine ID labels [1], angular margin loss enhancement [7], and spectral-temporal feature fusion [8]. Despite these advances, real-world industrial environments often suffer from strong background noise, which masks subtle anomalous sound characteristics and degrades detection reliability.

In this paper, we address the noise issue using dual-channel recordings captured by near-field and far-field microphones. The key idea is to exploit the spatial disparity between the two microphones: the near-field recording has a higher signal-to-noise ratio (SNR) as it is closer to the target machine, while the far-field recording captures predominantly background noise. Based on this observation, we propose a noise-aware spectral subtraction method that directly uses the far-field signal as a dynamic noise reference, eliminating the need for separate noise estimation. This is particularly advantageous in non-stationary factory environments where noise characteristics change rapidly.

A Conformer-based encoder is then employed for robust feature learning. The Conformer architecture combines convolutional modules and self-attention mechanisms, making it effective at capturing both local and global dependencies in the time-frequency domain. Under a multi-task self-supervised framework, we jointly train a machine-type classifier and a machine-attribute classifier, together with a contrastive loss, to learn discriminative embeddings. Finally, a kNN-based anomaly detector with SMOTE and domain-aware modeling is used to compute anomaly scores. Experimental results on the DCASE2026 Task 2 dataset demonstrate the effectiveness of the proposed method.

2. PROPOSED METHOD

2.1. Noise-aware Spectral Subtraction

Given a two-channel audio recording captured by microphones placed both near to and far from the target machine, let $\mathbf{X}_1(f,t)$ and $\mathbf{X}_2(f,t)$ denote the short-time Fourier transform (STFT) of the near-field and far-field signals, respectively, where f indexes the frequency bin and t the time frame. Since the two microphones are placed at different distances from the target machine, the near-field recording typically has a higher signal-to-noise ratio (SNR), while the far-field recording contains a stronger background noise component. To exploit this disparity, we propose a noise-aware spectral subtraction that directly uses the far-field signal as an estimate of the noise. The enhanced power spectrogram of the near-field signal is obtained as:

$$\hat{\mathbf{P}}(f,t) = max(|\mathbf{X}_1(f,t)|^2 - \alpha \cdot |\mathbf{X}_2(f,t)|^2, \beta \cdot |\mathbf{X}_1(f,t)|^2) \quad (1)$$

where α is an over-subtraction factor that controls the aggressiveness of noise suppression, and β is a spectral floor parameter used to avoid negative values and musical noise artifacts. In our experiments, $\alpha = 1.2, \beta = 0.01$.

In contrast to conventional single-channel spectral subtraction that requires a separate noise estimation stage, the proposed method leverages the far-field microphone as an always-available noise reference, which is particularly advantageous in non-stationary factory environments where noise characteristics can change rapidly.

After spectral subtraction, the enhanced spectrogram is converted to a log-Mel spectrogram by applying a Mel filter bank followed by a logarithmic compression, yielding a feature representation $\mathbf{M} \in \mathbb{R}^{D \times T}$, where D is the number of Mel bands and T the time frames.

2.2. Conformer Based Feature Learning

A Conformer-based encoder is used as the backbone for audio feature learning. The Conformer architecture combines convolutional modules and self-attention mechanisms, making it effective at capturing both local and global dependencies in the time-frequency domain. Given the input feature $\mathbf{M} \in \mathbb{R}^{D \times T}$, the Conformer outputs

Table 1: Results on DCASE2026 Task 2 development datasets

Method	Fan		Gearbox(Emu)		Bearing(Emu)		Slider(Emu)		ToyCar(Emu)		ToyCar		Valve(Emu)		hmean
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	
AE-MSE	64.10	60.42	52.54	52.29	63.61	53.97	57.77	50.36	68.06	53.47	65.23	58.25	56.55	50.20	57.65
AE-MAHALA	60.95	59.85	54.20	53.55	59.01	52.94	56.15	50.38	65.41	55.89	56.75	54.03	68.26	55.08	56.67
the proposed	85.83	74.74	89.71	81.58	63.69	59.63	63.95	52.74	70.44	54.74	66.58	55.37	61.87	52.63	66.68

a latent representation $\mathbf{Z} \in \mathbb{R}^{C \times D' \times T'}$, where C is the channel dimension, and D' and T' are the downsampled frequency and time dimensions, respectively. To obtain a fixed-dimensional embedding for each audio sample, we flatten the temporal and frequency axes by attentive pooling and then pass it through a linear projection layer, producing a compact embedding vector $\mathbf{e} \in \mathbb{R}^{L}$.

For self-supervised feature learning, this embedding is further fed to two auxiliary classification heads: a machine-type classifier and a machine-attribute classifier (e.g., speed or load level). Both heads are trained jointly using cross-entropy loss on the normal training data. Furthermore, a contrastive loss [1] is computed on the embedding in terms of machine-type labels to promote more compact feature space. The multi-task learning objective encourages the embedding space to capture discriminative characteristics of different machines and operating conditions.

2.3. kNN Based Anomaly Detector

For unsupervised anomaly detection, a k-nearest neighbors (kNN) approach is applied on the embeddings of all training normal samples, which form a reference memory bank. To tackle the domain shift issue, the SMOTE technique is employed to over-sample the embedding vectors of the target domain and domain-aware kNN models are trained separately. At test time, for a given test recording, the anomaly score is defined as the minimum Euclidean distance to these two kNN models.

3. EXPERIMENTS

3.1. Experimental Settings

Experiments were conducted on the DCASE2026 Task 2 development dataset [2, 3] that consists of seven machine types: Fan, Gearbox(Emu), Bearing(Emu), Slider(Emu), ToyCar(Emu), ToyCar, and Valve(Emu). Each audio recording is a two channel signal by two microphones with different distances to the target machine. The machine types marked with "Emu" are emulated two-channel recordings, whereas the remaining two are real two-channel recordings captured with two synchronized microphones. All audio recordings are padded to 12s with a sampling rate of 16kHz.

For audio preprocessing, 128-dimensional log-Mel energies are extracted after spectral subtraction and the input feature dimension is 512×3765 with a frame length of 1024 and a hop length of 512. The Conformer is trained for 20 epochs with a learning rate of 0.0001 and a batch size of 64. Mixup is applied on the input to generate inter-domain samples. Performance is evaluated using the official metrics of the challenge: AUC and pAUC.

3.2. Results

Our method achieves the highest harmonic mean AUC (66.68%) and pAUC (55.37%) on the DCASE2026 Task 2 dataset, outperforming all baselines. Notably, it attains 85.83% AUC on *Fan* and 89.71% on *Gearbox(Emu)*. The results validate the effectiveness of the proposed approach for noisy industrial anomaly detection.

4. CONCLUSION

This paper proposed a noise-aware spectral subtraction method using far-field signals as noise reference, combined with a Conformer-based encoder for robust feature learning under a self-supervised multi-task framework. A kNN-based detector with SMOTE and domain-aware modeling is used for anomaly detection. Experiments on the DCASE2026 Task 2 dataset show that the proposed method outperforms baseline systems, demonstrating its effectiveness for anomalous sound detection in noisy industrial environments.

5. REFERENCES

[1] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine id based contrastive learning pretraining," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[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] 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.

[5] Y. Wang, Q. Zhang, W. Zhang, and Y. Zhang, "A lightweight framework for unsupervised anomalous sound detection based on selective learning of time-frequency domain features," *Applied Acoustics*, vol. 228, p. 110308, 2025.

[6] J. Yang, "A two stage fusion anomaly detection approach for task2," in *DCASE 2025 Challenge Technical Report*, 2025.

[7] S. Choi and J.-W. Choi, "Noisy-arcmix: Additive noisy angular margin loss combined with mixup for anomalous sound

detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1–5.

[8] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 816–820.

[9] Y. Kawaguchi *et al.*, "Description and discussion on dcase 2022 challenge task 2: unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE)*, 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," in *Proceedings of 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.