ANOMALOUS SOUND DETECTION BASED ON NOISE REFERENCE ENHANCEMENT AND ADAPTIVE SCORE FUSION

Technical Report

Kaixing Ding, Da Huang, Aoyu Liu

Quectel Wireless Solutions Co., Ltd
Hefei, China
{kasen.ding, dane.huang, lay.liu}@quectel.com

ABSTRACT

This technical report presents the competition system designed by our team for the DCASE 2026 Anomalous Sound Detection (ASD) Challenge. We propose an anomalous sound detection system based on noise reference enhancement and adaptive score fusion. The system first estimates the noise power spectrum using the reference channel, computes the frequency-domain gain combined with the spectrum of the target channel, and obtains the enhanced audio. Then, complementary acoustic features are extracted, and a regularized statistical distance model is constructed based on normal samples. The multi-feature scores are adaptively fused according to the domain distance between the feature distributions of the source domain and target domain, and geometric product is utilized to impose consistency constraints on anomaly scores, so that samples judged as anomalous from multiple perspectives obtain higher anomaly scores. The system only requires normal audio clips from the development and evaluation datasets, with no need for synthetic anomalous samples. Test results on 7 types of development datasets show that the harmonic mean of AUC and pAUC of the proposed system reaches 0.6544, which surpasses the official baseline, and can meet the requirements of unsupervised anomaly detection in noisy industrial scenarios.

***Index Terms*—**Anomalous Sound Detection; Noise Reference Enhancement; Adaptive Score Fusion; Unsupervised

1. INTRODUCTION

Sound-based Anomalous Sound Detection (ASD) has attracted extensive attention from academia and industry in recent years due to its advantages of non-contact deployment, low cost and easy scalability[1]. Task 2 of the DCASE Challenge [2, 3, 4] aims to identify anomalous acoustic events for industrial equipment monitoring using only normal samples for training without prior knowledge of fault samples. On the basis of previous years, Task 2 of 2026 [5, 6, 7, 8] further introduces noise awareness and provides dual-channel stereo recorded data. The two channels are collected synchronously but have significant differences in Signal-to-Noise Ratio (SNR) and spectral characteristics, requiring improved ASD performance in noisy environments. This setting simulates the distribution shift caused by changes in equipment operating conditions in real industrial scenarios, and requires the detection model to effectively identify anomalies in noisy factory environments.

In response to the core changes of DCASE 2026 Task 2, this paper proposes an industrial anomalous sound detection system based on noise reference enhancement and adaptive score fusion. Firstly, the system estimates environmental noise based on the information differences between near-field and far-field channels, calculates the frequency-domain gain according to the power spectra of the target channel and reference channel noise, and applies the frequency-domain gain to the target channel spectrum to obtain the enhanced equipment sound. Multiple types of complementary acoustic features are extracted from the enhanced audio, and a regularized statistical distance model is established based on the normal training samples. The domain distance is quantified by the distribution distance between the source and target domains of the training samples in the feature space, and the fusion weights are adaptively determined according to the domain distance. Finally, the scores under different regularization strengths are fused, and consistency constraints are formed through geometric product, so that samples considered anomalous from multiple perspectives obtain higher anomaly scores, improving the effect of anomaly detection.

2. METHODOLOGY

2.1. Dual-Channel Noise Reference Enhancement

In our system, environmental noise is first estimated using the information differences between near-field and far-field channels. Generally, the near-field channel has a higher SNR relative to the target equipment, while the far-field channel contains more environmental noise and non-target sound sources. Therefore, the near-field channel is used as the target channel, and the far-field channel is used as the reference channel for estimating the noise components of the target channel. The frequency-domain gain $G(\tau, f)$ is constructed based on the power spectrum estimation of the target channel and reference noise, where τ denotes the time frame and f denotes the frequency index.

$$G(\tau, f) = \sqrt{\frac{\max(|X_1(\tau, f)|^2 - \alpha|X_2(\tau, f)|^2, 0)}{|X_1(\tau, f)|^2 + \varepsilon}}$$

Where $X_1(\tau, f)$ and $X_2(\tau, f)$ are the power spectra of the target channel and reference noise respectively; ε is a stability term to avoid division by zero; $\alpha \in [0,1]$ is the noise reduction coefficient, set to 0.5 in this case. Applying the gain to the target channel spectrum yields the enhanced target signal.

2.2. Extraction of Complementary Acoustic Features

After the enhanced audio is input into the feature extraction module, multiple types of acoustic features can be extracted in parallel to form a feature pool $F = \{F_1, F_2, \dots, F_n\}$. Not all features must be used simultaneously; the system adaptively selects one or more features according to the degree of domain shift.

1. BEATs [9] embedding features. We use a pre-trained BEATs encoder to extract global embeddings of the input waveform, capturing the global representation of audio and providing semantic-level or structural-level sound representations.
2. Sub-band time-frequency statistical features. For the Log-Mel spectrogram of the enhanced audio, sub-bands are divided by frequency, and statistics such as temporal mean, standard deviation, extreme values, band correlation and temporal autocorrelation are calculated to capture local spectral energy changes and harmonic stability.
3. Spectral shape features. Parameters such as spectral contrast, spectral centroid and spectral roll-off are adopted to describe the difference between harmonic peaks and background energy, suitable for anomalies such as air leakage, friction and broadband noise.
4. Multi-layer embeddings and multi-scale time-frequency features. For the BEAT model with multi-layer encoders, features can be extracted from shallow, middle and deep layers, and fixed-length vectors can be formed through PCA or other dimensionality reduction methods. Multi-scale time-frequency features such as constant-Q transform can also be extracted.

2.3. Regularized Statistical Distance Model

For the *j*-th feature $F_j(x)$ of the sample *x* to be detected, the mean vector μ_j and covariance matrix Σ_j are estimated based on normal training samples. To avoid singularity or instability of the covariance matrix under high-dimensional and small-sample conditions, a regularization term $\Sigma_{j,\lambda}$ is introduced:

$$\Sigma_{j,\lambda} = \Sigma_j + \lambda I$$

Where λ is the regularization strength and I is the identity matrix. The regularized statistical distance $d_{j,\lambda}(x)$ of $F_j(x)$ relative to the normal distribution is calculated as:

$$d_{j,\lambda}(x) = \left(F_j(x) - \mu_j\right)^T \Sigma_{j,\lambda}^{-1} \left(F_j(x) - \mu_j\right)$$

Due to domain shift between the source and target domains, directly using the global statistical distance may misclassify normal target-domain samples as anomalous. To this end, the system adopts local relative scoring, so that the sample to be detected is compared with the most extreme normal samples in the adjacent normal region instead of the global average normal sample. The *k* nearest neighbor samples $N_k(x)$ of the sample *x* to be detected are searched in the normal training samples, and the maximum normal distance of the neighbor samples is used as the local reference:

$$s_{j,\lambda}^{rel}(x) = \frac{d_{j,\lambda}(x)}{\max\limits_{u \in N_k(x)} d_{j,\lambda}(u) + \varepsilon}$$

When the sample to be detected does not exceed the maximum normal distance of normal samples in its neighborhood, its anomaly score will be reduced; when the sample to be detected is more extreme than all normal samples in its neighborhood, its anomaly score will be increased.

2.4. Domain Distance Adaptive Score Fusion

To adapt to the degree of domain shift of different equipment or working conditions, we explicitly define the domain-shift level according to the distance between the source-domain and target-domain feature centers. Let $F_b(x)$ denote the BEATs embedding of sample x, and let D_s and D_t denote the normal training samples from the source and target domains, respectively. The BEATs-based domain distance is computed as:

$$g = \left\| \frac{1}{|D_s|} \sum_{x \in D_s} F_b(x) - \frac{1}{|D_t|} \sum_{x \in D_t} F_b(x) \right\|_2$$

The BEATs embedding space is used because it provides a compact representation of global acoustic patterns and is less sensitive to local spectral fluctuations than low-level statistical features. Therefore, it is more suitable for estimating the overall mismatch between the source and target domains.

Based on g, the system divides the domain shift into four levels, as shown in Table1. When $g \le 0.8$, the source and target domains are considered very close. In this case, the target-domain normal samples are still located near the source-domain normal manifold, so all complementary features are enabled. When $0.8 < g \le 1.0$, the shift is mild, but features sensitive to recording conditions may introduce unstable scores; therefore, these features are down-weighted or disabled. When $1.0 < g \le 1.4$, the shift is moderate, and low-level spectral statistics may begin to reflect domain mismatch rather than abnormality; therefore, more robust feature groups are prioritized. When $g > 1.4$, the shift is large, and the system mainly relies on embedding-based anomaly scores, because low-level statistical features may primarily capture recording-condition differences.

Table1: The values of g and the corresponding fusion strategies

Values of g	Domain-shift level	Adaptive fusion strategy
$g \le 0.8$	Small	Enable all complementary features
$0.8 < g \le 1.0$	Mild	Down-weight or disable condition-sensitive features
$1.0 < g \le 1.4$	Moderate	Prioritize robust feature groups
$g > 1.4$	Large	Mainly rely on embedding-based scores

For each feature group, the raw anomaly score is first converted into a rank-normalized score $R_j(x)$ to reduce the scale difference between different scoring functions. The final anomaly score is then obtained by geometric fusion:

$$S(x) = \prod_{j=1}^{J} R_j(x)^{w_j(g)}, \sum_{j=1}^{J} w_j(g) = 1$$

Where $w_j(g)$ is the weight of the *j*-th feature determined by the domain-shift level, and J is the number of feature categories

participating in fusion. This design allows the system to use richer multi-view information when the domains are close, while falling back to more robust embedding-based scores when the domain mismatch becomes severe. A larger $S(x)$ indicates a higher probability that the sound to be detected is anomalous. The system can directly output $S(x)$, or output normal/anomalous judgment and alarm signals after comparing $S(x)$ with a threshold.

3. EXPERIMENTS

We evaluated the system on the development dataset of DCASE 2026 Task 2, which contains sound samples of source and target domains under various equipment types. The BEATs model directly uses pre-trained weights without fine-tuning. The system is compared with the baseline systems AE-MSE and AE-MAHALA of DCASE 2026 Task 2. The results are shown in Table2. Our system achieves a harmonic mean of AUC and pAUC of 0.6544 on the development dataset, outperforming AE-MAHALA (0.5766) and AE-MSE (0.5666).

Table2: AUCs and pAUCs per machine type obtained on the development dataset

Machine	Metric	Baseline(mahala)	Baseline(mse)	Our system
ToyCarEmu	AUC-S	0.6949	**0.6962**	0.6400
	AUC-T	**0.6662**	0.6120	0.6244
	pAUC	0.5347	**0.5589**	0.5553
ToyCar	AUC-S	**0.7728**	0.7562	0.5991
	AUC-T	0.5317	0.3787	**0.7866**
	pAUC	**0.5825**	0.5403	0.5561
bearingEmu	AUC-S	**0.6592**	0.6234	0.6462
	AUC-T	0.6228	0.5956	**0.6758**
	pAUC	**0.6042**	0.5985	0.5905
fan	AUC-S	0.6000	0.6145	**0.6402**
	AUC-T	0.4509	0.4694	**0.6474**
	pAUC	0.5229	0.5333	**0.5800**
gearboxEmu	AUC-S	0.7448	0.6823	**0.7538**
	AUC-T	0.5274	0.4978	**0.6900**
	pAUC	0.5397	0.5294	**0.6332**
sliderEmu	AUC-S	0.6636	0.6725	**0.7766**
	AUC-T	0.4918	0.4505	**0.5172**
	pAUC	0.5036	0.5038	**0.5358**
valveEmu	AUC-S	0.5660	0.6774	**0.9248**
	AUC-T	0.5650	0.6878	**0.9324**
	pAUC	0.5020	0.5508	**0.8063**
H.mean	AUC-S	0.6646	0.6718	**0.6974**
	AUC-T	0.5424	0.5085	**0.6762**
	pAUC	0.5391	0.5437	**0.5982**
	H.mean	0.5766	0.5666	**0.6544**

AUC-S and AUC-T are the AUC of the source and target domains, respectively.

4. CONCLUSION

For DCASE 2026 Task 2, a sound anomaly detection system based on noise reference enhancement and adaptive score fusion is proposed. Leveraging the dual-channel data provided by the challenge, the system performs frequency-domain noise suppression by using the far-field channel as the noise reference and the near-field channel as the target channel, which effectively improves the sound quality of target equipment in noisy environments. A multi-view representation framework is constructed by extracting complementary acoustic features including BEATs embeddings, sub-band time-frequency statistics, and spectral shape features. A regularized statistical distance model combined with a local relative scoring mechanism is adopted to alleviate detection bias caused by domain shift. Fusion weights of multiple features are adaptively adjusted according to the domain distance between feature distributions of the source domain and target domain. Finally, a consistency constraint on anomaly scores is imposed via geometric product to strengthen high-confidence anomaly decisions. Experimental results demonstrate that the proposed system can be trained only with normal samples and requires no anomalous samples. On the seven types of development datasets, it achieves a harmonic mean of AUC and pAUC of 0.6544, outperforming baseline models based on reconstruction algorithms. This validates the effectiveness and robustness of the proposed method for anomalous sound detection in complex industrial noise scenarios and under domain shift conditions.

5. REFERENCES

[1] K.Wilkinghoff, "Self-supervised learning for anomalous sound detection," *in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 276–280.

[2] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2305.07828*, 2023.

[3] 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 2024 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2406.07250*, 2024.

[4] 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," *In arXiv e-prints: 2506.10097*, 2025.

[5] 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.

[6] 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)*, 1–5. Barcelona, Spain, November 2021.

[7] 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.

[8] 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), pages 191–195, 2023.

[9] 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," *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193.