ENHANCED UNSUPERVISED ANOMALOUS SOUND DETECTION VIA CONVTASNET-BASED SEPARATION AND CONDITIONAL AUTOENCODING

Technical Report

Chenjun Fu [1], Ronghuan Zhao[1],Qiang Wang[1],Hao Wu[1],Liang Zou[1]

[1] China University of Mining and Technology, Xuzhou, China
{fcj, ronghuanzhao,qiangwang,haowu,liangzou}@cumt.edu.cn

ABSTRACT

This report outlines our approach to first-shot unsupervised anomalous detection for machine condition monitoring, developed for DCASE 2025 Task 2. Given the constraint of only having normal operational data and the availability of clean target device sounds or background noise, our method focuses on leveraging audio separation and a self-supervised AutoEncoder (AE) for anomaly detection.

Key components of our approach include training an audio separation module to extract target sounds for effective denoising and data augmentation, encoding audio features via an AutoEncoder trained solely on normal data, and performing conditional modeling with attribute and domain labels 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 and its nearest neighbors in the training set; greater distances imply higher anomaly likelihood.

Our approach achieved notable performance on the development set,demonstrating is effectiveness. The AUC for the target domain was 64.1% and for the source domain was 60.8%. Addtitionally, the Partial AUC values (p=0.1) for the target and source domain was 55.6 % . 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, conditional Autoencoder, reconstruction loss, log-mel spectrogram

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 2025 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]. This year's iteration highlights a "first-hot" problem under both attribute-available and attribute-unavailable conditions. In practice, the heterogeneity of machine types poses greater challenges for utilizing sounds collected with trainable attribute labels. Consequently, the current task features:

1. An updated and expanded set of machine types for evaluation.
2. Provision of pure noise or pure machine sound data for specific machine types.
3. Absence of attribute labels for training certain machine types.

Based on the current task configuration, we observe that the provided isolated operational sounds and background noise samples offer advantageous conditions for constructing audio separation modules[2,3]. This enables effective denoising and signal enhancement of the original mixed recordings. To leverage this opportunity, we introduce a Conv-TasNet-based audio separation module, which employs a convolutional time-domain architecture designed to effectively capture local temporal structures and perform fine-grained separation of overlapping sound sources. Through processing raw audio with this denoising framework, Conv-TasNet significantly improves the signal-to-noise ratio (SNR) and enhances signal fidelity, facilitating morerobust downstream feature modeling and anomaly detection.

Following denoising, we employ a self-supervised AutoEncoder (AE) trained exclusively on normal operating data to learn compact latent representations of the purified audio. The AE is designed to reconstruct normal signals with minimal error, enabling it to implicitly model the distribution of normal acoustic patterns. Deviations from this distribution—quantified by reconstruction error or latent-space distance—serve as indicators of potential anomalies. These learned representations provide essential support for anomaly discrimination and offer a degree of domain invariance by abstracting machine-specific acoustic features.

The synergistic integration of audio separation and latent representation learning enables our system to accurately characterize machine state distributions in acoustically complex industrial environments. Consequently, this approach substantially enhances anomaly detection performance, particularly when encountering unseen machine types and operating under weakly-labeled or low-resource conditions.

2. METHODOLOGY

2.1 Audio Separation Strategy with Conv-TasNet

Given the task configuration of DCASE 2025 Task 2, where isolated operational machine sounds and background noise samples are provided, we identify a unique opportunity to build a high-performance audio separation module tailored for machine condition monitoring. These clean references enable the learning of explicit mappings between mixture signals and target components, thereby facilitating supervised denoising and signal decomposition.

To fully exploit this setting, we adopt Conv-TasNet[4], a time-domain audio separation network known for its effectiveness in modeling local temporal structures with high-

resolution detail. Conv-TasNet uses a learnable encoder-decoder framework with 1D convolutional filters and a temporal convolutional network (TCN) as the separation core. Unlike traditional spectrogram-based models, this architecture operates directly on raw waveforms, enabling precise recovery of fine-grained machine sound components while suppressing irrelevant background noise.

Let $X_{mix} \in R^N$ denote the observed mixture signal. The Conv-TastNet model decomposes it into estimated source $\hat{X} = \hat{X}_{mach} + \hat{X}_{noise}$, where each stream is optimized using scale- invariant source-to-noise ratio (SI-SNR) loss[5] against the clean references. This formulation allows the model to act as a preprocessing denoising stage, producing enhanced audio $\hat{X}_{mach}$ for downstream anomaly detection.

Learnable Econder: The input mixture waveform $X_{mix} \in R^N$ is first transformed into a latent representation using a 1D convolutional encoder:

$$X = \text{Encoder}(X_{mix}) \in R^{C \times T}$$

This encoder acts as a learnable filterbank that extracts local temporal structures while preserving the non-stationary and high-frequency characteristics of the original signal, outperforming traditional STFT-based methods[6].

Separation Module: The encoded features X are then passed through stacked Temporal Convolutional Networks (TCNs) to estimate source-specific masks:

$$\hat{M} = \text{Spearator}(X) \in R^{S \times C \times T}$$

where S is the number of sources (typically 2: machine sound and noise). The masks $\hat{M}$ indicate the activation patterns of each source in the latent space. The TCNs[7], with causal convolutions and residual connections, enable efficient long-range temporal modeling tailored to continuous and non-stationary industrial acoustic signals.

Decoder: Each masked latent representation is transformed back to the waveform domain using a 1D transposed convolutional decoder:

$$\hat{X}^{(S)} = \text{Dcoder}(\hat{M}^{(S)} \circ X), \ S=1,2$$

The reconstructed waveform $\hat{X}^{(1)}$ is treated as the denoised machine sound and used as input for subsequent feature extraction.

To train this module, we leverage the provided isolated source data with a supervised loss based on Scale-Invariant Signal-to-Noise Ratio (SI-SNR):

$$L_{sep} = -\text{SI-SNR}(\hat{X}_{mach}, X_{mach}) - \text{SI-SNR}(\hat{X}_{noise}, X_{noise})$$

With this learning strategy, the model not only extracts the target machine sounds effectively but also suppresses complex environmental noise and non-structured background information, significantly improving the overall signal-to-noise ratio (SNR)[8].

Furthermore, due to Conv-TasNet's strong capability in modeling local temporal patterns and short-range dependencies, the separated signals retain critical semantic features of the machine operation. This results in cleaner and more robust inputs for feature encoders (e.g., autoencoders), substantially enhancing anomaly detection performance, particularly under unseen machine types and domain-shifted conditions.

2.2 Conditional Autoencoder

Autoencoder (AE) detects anomalous sounds based on reconstruction loss[9] . Specifically, the encoder component maps the input feature vector to a low-dimensional latent representation, and the decoder component attempts to reconstruct the original input signal from this latent representation. The reconstruction loss is defined as the difference between the original input feature vector and the output vector produced by the AE . For samples not present in the training set (i.e., anomaly samples), the reconstruction loss of the AE will increase significantly, allowing them to be identified as abnormal.

In terms of data processing, we convert the STFT into log-Mel spectrogram for better feature representation[10]. To convert the filtered STFT to a log-Mel spectrogram, we apply a Mel filter bank $M(f)$ and take the logarithm:

$$S_{Mel}(m,t) = \log\left(\sum_{f=0}^{F-1} |X(t,f)|^2 \cdot M(m,f)\right)$$

Where：

- $S_{\text{Mel}}(m,t)$ is the log-Mel spectrogram.
- $M(m,f)$ is the Mel filter for the m-th Mel frequency bin.

To enhance anomaly detection performance in specific contexts, we adopt a Conditional Autoencoder (cAE) framework. Machine-related information (e.g., machine ID or type) is encoded into a condition vector *c*, which is concatenated with the log-Mel spectrogram S_{Mel} and fed into the autoencoder to guide reconstruction[11]. This allows the model to better adapt to different machine conditions. Formally, the reconstruction is defined as:

$$\hat{S} = f_{AE}(S_{Mel}, c)$$

Where f_{AE} denotes the reconstruction function of the conditional autoencoder. The reconstruction loss is calculated as the mean squared error (MSE) between the input and its reconstruction:

$$L_{recon} = \| S_{\text{Mel}} - S \|^2$$

During training, we compute the reconstruction errors for all samples in the source domain and store them as a reference distribution[12]. In the testing phase, we similarly reconstruct the input and compute its reconstruction error vector $e_{test} = \|S_{Mel} - \hat{S}\|^2$. To determine the anomaly score, we employ a K-Nearest Neighbors (KNN) based approach[13]. Specifically, we compute the average Euclidean distance between the test sample and its *K* nearest neighbors from the training set in the error space:

$$D_{\text{KNN}} = \frac{1}{K} \sum_{i=1}^{K} \| e_{test} - e_i^{train} \|_2$$

Where e_i^{train} denotes the reconstruction error vectors of the *K* nearest training samples. To improve robustness across domains, we compute this score separately for the source and target domains, and take the minimum of the two as the final anomaly score:

$$D_{\text{final}} = \min\left\{D_{\text{KNN}}^{\text{source}}, D_{\text{KNN}}^{\text{target}}\right\}$$

To further alleviate discrepancies in score distributions between machines, we apply domain-wise score normalization:

$$L_{\text{anomaly}} = \frac{D_{\text{final}} - \mu_d}{\sigma_d}$$

Where μ_d and σ_d are the mean and standard deviation of KNN scores within the respective domain *d* (source or target).

This method combines conditional modeling, local error structure via KNN, and cross-domain normalization to achieve robust anomaly detection with improved generalization across unseen machine conditions.

Table 1: DCASE 2025 Task 2 experimental results on development dataset (%). The value in the row "Total Score" represents the harmonic mean of the AUC and pAUC scores over all the machine types and domains.

		Baseline (MSE)	Baseline (MAHALA)	Our system
ToyCar	AUC(source)	71.05%	73.17%	96.6%
	AUC(target)	53.52%	50.91%	78.2%
	pAUC	49.7%	49.05%	59.8%
ToyTrain	AUC(source)	61.76%	50.87%	58.0%
	AUC(target)	56.46%	46.15%	64.4%
	pAUC	50.19%	48.32%	51.1%
bearing	AUC(source)	66.53%	63.63%	54.4%
	AUC(target)	53.15%	59.03%	48.9%
	pAUC	61.12%	61.86%	50.6%
fan	AUC(source)	70.96%	77.99%	52.2%
	AUC(target)	38.75%	38.56%	60.0%
	pAUC	49.46%	50.82%	50.2%
gearbox	AUC(source)	64.8%	73.26%	80.3%
	AUC(target)	50.49%	51.61%	73.9%
	pAUC	52.49%	55.07%	62.6%
slider	AUC(source)	70.1%	73.79%	78.1%
	AUC(target)	48.77%	50.27%	59.5%
	pAUC	52.32%	53.61%	56.2%
valve	AUC(source)	63.53%	56.22%	74.8%
	AUC(target)	67.18%	61.0%	81.9%
	pAUC	57.35%	52.53%	55.7%
All	AUC(source)	66.77%	65.51%	67.5%
	AUC(target)	51.39%	50.05%	64.8%
	pAUC	52.94%	52.72%	54.8%

Our method begins with a separation network that performs denoising and disentanglement on raw audio signals. This pre-processing step aims to suppress background noise and isolate machine-related sound components, thereby improving the quality of features used for anomaly detection. The separated machine sounds are then treated as augmented data and used as inputs to a convolutional autoencoder (AE).

We utilize 128-dimensional log-Mel spectrogram features extracted from the audio as the input to the AE. The convolutional AE is designed to learn robust representations and reconstruct clean signals conditioned on machine attributes. The training is conducted with a batch size of 256, using the Adam optimizer with a learning rate of 0.001.

In scenarios where machine labels are missing or unavailable, we adopt attribute classification or clustering-based strategies to generate pseudo-labels for training the conditional AE. The rest of the training and scoring procedures follow the KNN-based anomaly detection approach described previously.

3. RESULT

Table 1 presents the results of our system. Compared to the baseline. As a key enhancement, we first apply an audio separation strategy using Conv-TasNet to isolate machine-relevant sounds from background noise. The separated machine sounds are treated as augmented data and used as input to the downstream anomaly detection models. This data augmentation improves the robustness and generalization of the system, especially under challenging acoustic environments.

Compared to the baseline, our improved Conditional AE[14], along with the score normalization scheme for KNN-based on anomaly scores across source and target domains, showed of slightly lower performance in the source domain but significantly better performance in the target domain. Overall, our Conditional AE-based anomaly sound detection model demonstrated notable improvements over the baseline, enhancing the detection performance.

We submitted four systems for Task 2 of the DCASE 2025 Challenge, all of which have the same processing pipeline except:

1. The conditional Autoencoder that uses conditional inputs for improved anomaly detection
2. The Autoencoder without conditional inputs.
3. A system with 256-sized log-Mel features for higher resolution analysis.
4. A Fully Connected Network mimicking the baseline structure with fully connected layers instead of convolutional layers.

4. CONCLUSION

In this technical report, we present our submission for Task 2 of the DCASE 2025 Challenge. Our proposed system is based on a Conditional Convolutional Autoencoder (AE) architecture, enhanced by a prior audio separation stage. Specifically, we introduce a Conv-TasNet-based separation module that processes the original machine sound recordings to isolate machine-relevant components and suppress background noise. The separated machine sounds are then used as augmented training data, effectively improving the quality and diversity of input features for the downstream AE.

To extract meaningful representations from the audio, we compute 128-dimensional log-Mel spectrograms from the separated signals. These are used as inputs to the convolutional AE, which is trained to reconstruct clean representations conditioned on machine attributes. This approach allows the model to better capture machine-specific behaviors and improve anomaly detection under varying operational conditions.

When attribute or domain labels are available, they are directly encoded and fed into the conditional AE. In cases where such metadata is missing, we apply classifier-based inference or clustering algorithms to estimate pseudo-labels for model training. During inference, we adopt a K-Nearest Neighbor (KNN) strategy in the reconstruction error space to compute anomaly scores, and apply domain-wise score normalization to reduce distribution bias across different machine types.

Our experimental results on the development set demonstrate that the integration of audio separation and conditional reconstruction significantly improves anomaly detection performance compared to baseline methods.

5. REFERENCES

[1] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei 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.

[2] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda, and Shoichiro 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.

[3] Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei 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] Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation[J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 27(8): 1256-1266.

[5] Fu X S, Yao S Y, Xu J T, et al. Study on high signal-to-noise ratio (SNR) silicon pn junction photodetector[J]. Optica Applicata, 2006, 36.

[6] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 26, no. 1, pp. 1702–1726, Oct. 2018.

[7] Farha Y A, Gall J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3575-3584.

[8] Tandra R, Sahai A. SNR walls for signal detection[J]. IEEE Journal of selected topics in Signal Processing, 2008, 2(1): 4-17.

[9] P. Li, Y. Pei, and J. Li, “A comprehensive survey on design and application of autoencoder in deep learning,” Applied Soft Computing, vol. 138, p. 110176, 2023.

[10] H. Tao, P. Wang, Y. Chen, V. Stojanovic, and H. Yang, “An unsupervised fault diagnosis method for rolling bearing using stft and generative neural networks,” Journal of the Franklin Institute, vol. 357, no. 11, pp. 7286–7307, 2020.

[11] Imai S. Cepstral analysis synthesis on the mel frequency scale[C]//ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1983, 8: 93-96.

[12] L. Yang and Z. Zhang, “A conditional convolutional autoencoder-based method for monitoring wind turbine blade breakages,” IEEE transactions on industrial informatics , vol. 17, no. 9, pp. 6390–6398, 2020.

[13] Guo G, Wang H, Bell D, et al. KNN model-based approach in classification[C]//On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer Berlin Heidelberg, 2003: 986-996.

[14] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro 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.