# SJTU-AITHU SYSTEM FOR DCASE 2025 ANOMALOUS SOUND DETECTION CHALLENGE

Technical Report

*Xinhu Zheng*[1], *Anbai Jiang*[2], *Bing Han*[1], *Shuwei Zhang*[3]
*Wei-Qiang Zhang*[3], *Xie Chen*[1], *Cheng Lu*[4], *Pingyi Fan*[2], *Jia Liu*[2,3], *Yanmin Qian*[1]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Tsinghua University, Beijing, China
[3] Huakong AI Plus Company Limited, Beijing, China
[4] North China Electric Power University, Beijing, China
Email: zhengxh24@sjtu.edu.cn, jab22@mails.tsinghua.edu.cn

## ABSTRACT

This report presents our solutions for DCASE 2025 Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. In this domain, pre-trained models have demonstrated considerable potential, particularly in handling domain shifts. We develop our systems based on BEATs and the EAT family and explore various training strategies to enhance performance. Sub-center loss and noise-aware training are employed to improve system performance. By fusing various models and methods, we achieve an hmean of 69.12% on the development dataset.

***Index Terms*—** DCASE Challenge, anomaly detection, sound, pre-trained model, noise-aware training

## 1. INTRODUCTION

In the realm of industrial automation, the ability to detect anomaly sounds remains vital for ensuring operational reliability and preventing potential failures. The DCASE 2025 Challenge Task 2 [1, 2, 3, 4], namely First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring, continues to focus on identifying anomalies in sounds from specific machine types. This year's challenge introduces additional complexity by expanding the dataset to include both clean data and pure noise samples, which further tests the robustness of algorithms in distinguishing genuine anomalies from normal operational noise.

The complexity of this task lies in accurately distinguishing between normal operational noise and genuine anomalies, requiring algorithms capable of learning from diverse acoustic patterns. In practical production environments, the diversity of equipment types, complex surroundings, and challenges with sound data collection make it difficult to develop systems that can accurately identify and classify abnormal sounds across different devices and environments. The main challenges can be summarized as follows:

- **Data scarcity for training**. While this year's dataset includes more samples, real industrial scenarios still face the fundamental issue of limited data for model training. Models must still overcome the challenge of learning from relatively scarce examples of normal operations without anomalies.
- **Domain shifts**. The complexity of industrial production environments, varied background noises, and differences in recording equipment continue to cause distribution differences in audio data. The additional clean data in this year's dataset may help mitigate but not eliminate these domain shift issues.
- **Incomplete training labels**. The data collection process still faces the problem that not all machine types have available attribute labels. Models must maintain good generalization performance despite these limitations.

Following our previous works [5, 6, 7, 8], we continue to leverage pre-trained models to provide necessary generalization capability across different machines. This year, due to changes in competition rules that permit the use of all available data from DCASE 2020 to DCASE 2025 for training, we find that certain methods, such as LoRA [9] tuning and SMOTE [10], are no longer effective when the data size scales up. Consequently, we choose not to employ these approaches. Instead, we introduce noise-aware training and sub-center loss [11] to enhance model robustness and address the issue of missing labels. All submitted systems are ensemble systems where single model scores are combined, with our best system achieving a harmonic mean of 69.12% on the development set.

The structure of the paper is organized as follows. Section 2 introduces the pre-trained models and the additional strategies. Section 3 gives an overview of all the submitted systems. Section 4 presents the detection results.

## 2. METHODS

### 2.1. BEATs

BEATs [12], short for Bidirectional Encoder representation from Audio Transformers, has demonstrated superior performance compared to alternative pre-trained audio models. This self-supervised learning framework employs an iterative optimization process between its acoustic tokenizer and audio self-supervised learning (SSL) model components. The architecture generates rich semantical discrete labels that effectively capture audio representations, which is particularly beneficial for our classification objectives and anomaly detection. We utilize the BEATs-iter3 version, which was pre-trained on AudioSet [13] and contains 90M parameters.

For model adaptation, we perform attribute-based fine-tuning across all machine types for DCASE 2020 to DCASE 2025. The input processing pipeline standardizes audio segments to 10-second durations, followed by log-mel spectrogram conversion using 25 ms frames, 10 ms frame shifts, and 128 mel bins. To enhance ro-

bustness, we apply SpecAug [14] with a maximum masking length of 80 on both the time dimension and the frequency dimension. The model architecture incorporates an attentive statistics pooling layer for frame-to-utterance embedding aggregation from ECAPA-TDNN [15]. Two dense layers are appended to predict the logits. Our classification approach dynamically adapts to label availability: for samples with strong attribute annotations (DCASE 2022-2025 data), we use these attribute labels as classification targets; when only weak or no attribute labels exist (DCASE 2020-2021 data), we fall back to section labels as the classification criterion. This hybrid strategy maximizes the utilization of all available labels across different datasets. Training employs ArcFace [16] loss over 30,000 steps using AdamW [17] optimization, with 8-step gradient accumulation, 360 warm-up steps, and batch size 32.

The anomaly detection system computes cosine distances between embeddings using a similarity score with a 1-nearest neighbor approach, where the minimum distance to any training sample serves as the anomaly score for each test instance.

### 2.2. Other SSL Models

In addition to BEATs, we also investigate the use of EAT [18] and a self-developed SSL model. The self-developed model adopts short-time fourier transform (STFT) as the input and models the sub-band in a teacher-student framework. It is trained on 17k hours of audio from Audioset [19], Freesound[1], MTG-Jamendo [20] and Music4all [21]. The self-developed model will be introduced in detail in an upcoming reseach paper. Both models are fine-tuned on all six DCASE datasets. During detection, we extract the [CLS] embedding and conduct the identical anomaly detection pipeline with BEATs.

### 2.3. Noise-Aware Training

To improve model robustness against environmental noise, we implement noise-aware training utilizing the provided pure noise samples in the DCASE 2025 dataset. During training, each audio sample has a 50% probability of being mixed with randomly selected noise at varying SNRs. Specifically, for each potentially corrupted sample, we:

- Randomly select one noise sample from the provided collection
- Randomly choose an SNR level from $\{5, 10, 15, 20\}$ dB
- Mix the original audio with noise at the selected SNR level
- Maintain the original label regardless of noise addition

This approach serves two key purposes: (1) it regularizes the model against overfitting to the original training data, and (2) better prepares the system for real-world conditions where machine sounds often coexist with environmental noise. The SNR range is selected to cover both challenging (5dB) and more moderate (20dB) noise conditions, representing realistic industrial scenarios. Notably, we only use the officially provided noise samples from the DCASE 2025 dataset, ensuring consistency with the evaluation environment.

[1] https://freesound.org/

### 2.4. Sub-Center Loss

We employ a Sub-Center ArcFace loss [11] to handle label ambiguity and improve feature discrimination. The loss function enhances traditional angular margin approaches by introducing multiple sub-centers for selected classes:

$$\mathcal{L} = -\log \frac{e^{s(\cos(\theta_y + m))}}{e^{s(\cos(\theta_y + m))} + \sum_{j \neq y} e^{s \cos \theta_j}} \tag{1}$$

where $\theta_y$ represents the angle between the embedding and its nearest sub-center for target class $y$, $s = 30$ is the scaling factor, and $m = 0.2$ is the angular margin. Key implementation details include:

- Sub-centers ($k = 16$) are only activated for: (1) all DCASE 2020-2021 machine types, and (2) DCASE 2024-2025 machine types without attribute labels
- Standard ArcFace (single center) is used for other cases
- During training, each sample automatically associates with its nearest sub-center
- The margin penalty helps create more discriminative feature spaces

This selective application of sub-centers provides two benefits: (1) robustness for poorly-labeled or attribute-missing samples through introducing extra representation, while (2) maintaining simpler discrimination boundaries for well-labeled classes.

## 3. SUBMITTED SYSTEMS

Our four submitted systems comprise ensembles derived from 13 systems, including one baseline system combining BEATs, EAT, and the self-developed model with only sub-center loss applied, and 12 systems based on BEATs, employing different method combinations and hyperparameters. For each system, we selected the top-3 performing checkpoints during training for the internal ensemble.

System 1 implements the fusion of all the systems based on BEATs. System 2 presents the fusion across all 13 systems. System 3 combines the baseline system with two systems that show particularly robust score distributions on the DCASE 2025 evaluation set. System 4 merges the baseline system with two top-performing systems based on quantitative metrics.

Models are ensembled by linearly combining the anomaly scores of different models, where the coefficients are attained by either grid search or Bayesian optimization. The Bayesian approach in system 1 automatically determines weights, whereas systems 3-4 use grid search to find coefficients that balance performance and robustness. This multi-strategy ensemble framework provides both comprehensive model averaging and targeted performance optimization.

## 4. EXPERIMENT RESULTS

The detection performance is evaluated using the standard metrics specified in the DCASE 2025 challenge: the Receiver Operating Characteristic (ROC) curve's Area Under Curve (AUC), partial AUC (pAUC) in the false positive rate range of 0-0.1, and their harmonic mean. For each machine type, we compute both source and target domain AUC scores along with pAUC values, then combine them through harmonic averaging according to the official evaluation baseline.

Table 1: Results of four submitted systems on the development set

| Machine | Metric | System 1 | System 2 | System 3 | System 4 |
|---|---|---|---|---|---|
| bearing | AUC_s | 68.26 | 66.76 | 65.74 | 66.22 |
| | AUC_t | 67.96 | 68.44 | 68.06 | 68.42 |
| | pAUC | 60.68 | 58.79 | 58.00 | 58.11 |
| | hmean | 65.44 | 64.38 | 63.63 | 63.93 |
| fan | AUC_s | 61.54 | 61.42 | 61.24 | 61.36 |
| | AUC_t | 62.40 | 62.50 | 62.56 | 62.06 |
| | pAUC | 55.42 | 55.89 | 55.42 | 55.79 |
| | hmean | 59.62 | 59.79 | 59.57 | 59.60 |
| gearbox | AUC_s | 80.96 | 82.92 | 83.10 | 82.90 |
| | AUC_t | 75.20 | 82.22 | 82.94 | 82.50 |
| | pAUC | 68.79 | 67.53 | 67.79 | 67.53 |
| | hmean | 74.65 | 76.86 | 77.24 | 76.94 |
| slider | AUC_s | 88.40 | 91.98 | 91.92 | 92.12 |
| | AUC_t | 76.22 | 77.82 | 77.30 | 77.92 |
| | pAUC | 60.63 | 59.95 | 59.68 | 60.05 |
| | hmean | 73.30 | 74.25 | 73.94 | 74.36 |
| ToyCar | AUC_s | 67.82 | 69.04 | 69.12 | 69.10 |
| | AUC_t | 63.22 | 63.26 | 62.76 | 63.20 |
| | pAUC | 51.79 | 53.16 | 52.84 | 53.11 |
| | hmean | 60.15 | 61.10 | 60.82 | 61.07 |
| ToyTrain | AUC_s | 75.76 | 76.92 | 76.50 | 77.22 |
| | AUC_t | 74.70 | 70.34 | 69.20 | 70.26 |
| | pAUC | 62.11 | 57.63 | 56.53 | 57.47 |
| | hmean | 70.28 | 67.31 | 66.35 | 67.29 |
| valve | AUC_s | 88.16 | 87.58 | 87.32 | 87.44 |
| | AUC_t | 93.92 | 94.06 | 93.60 | 94.04 |
| | pAUC | 78.11 | 84.11 | 84.63 | 84.11 |
| | hmean | 86.22 | 88.39 | 88.36 | 88.34 |
| hmean | AUC_s | 74.59 | 75.19 | 74.91 | 75.13 |
| | AUC_t | 72.16 | 72.70 | 72.36 | 72.63 |
| | pAUC | 61.53 | 61.17 | 60.75 | 61.02 |
| | hmean | 68.94 | **69.12** | 68.76 | 69.02 |

AUC_s and AUC_t are the AUC of the source and target domains, respectively.

Table 1 shows the detailed performances of four submitted systems. The AUC_s, AUC_t, pAUC, and the harmonic mean are calculated for each system and machine type. The best result on the DCASE 2025 development dataset is achieved by system 2, with a harmonic mean of 69.12%.

## 5. CONCLUSION

This paper presented the SJTU-AITHU system for DCASE 2025 Task 2 on first-shot unsupervised anomalous sound detection. Our approach leveraged BEATs and EAT pre-trained models enhanced with noise-aware training to improve robustness against environmental interference, while employing sub-center loss to address label missing and misalignment issues across different machine types and datasets. As a result, the proposed system achieved a best harmonic mean of 69.12% on the development set.

## 6. REFERENCES

[1] 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.

[2] 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.

[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] 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.

[5] Z. Lv, A. Jiang, B. Han, Y. Liang, Y. Qian, X. Chen, J. Liu, and P. Fan, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[6] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Thuee system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[7] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," in *Interspeech 2024*, 2024, pp. 107–111.

[8] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 969–974.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[11] A. Jiang, X. Zheng, B. Han, Y. Qiu, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Adaptive prototype learning for anomalous sound detection with partially known attributes," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[12] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[13] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[16] 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.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[18] 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.

[19] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[20] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. [Online]. Available: http://hdl.handle.net/10230/42015

[21] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues, *et al.*, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2020, pp. 399–404.