A CONSTRAINT-DRIVEN STATISTICAL PIPELINE FOR NOISE-AWARE ANOMALOUS SOUND DETECTION UNDER FREE-TIER COMPUTE

Technical Report

Mehdi Zarrouky

Independent Researcher, Casablanca, Morocco
zarrouky@gmail.com

ABSTRACT

We describe a fully statistical anomaly detection system submitted to DCASE 2026 Challenge Task 2 [3], developed under severe computational constraints (free-tier cloud notebooks, CPU only, no persistent compute environment). These constraints precluded the fine-tuning of large pre-trained acoustic models and shaped a methodology focused on parameter-free statistical inference rather than learned representations. The pipeline combines dual-channel exploitation (log-mel and constant-Q from both microphones and their difference), multi-scale temporal pooling, L2-normalized domain-aware Mahalanobis scoring with Ledoit–Wolf shrinkage [5], and rank-based ensemble. The system uses no neural networks and no external datasets. On development data, including ToyADMOS2 [1] and MIMII-derived machines [2], and following the first-shot anomaly detection paradigm [4], the system achieves a harmonic mean of 60.76 across AUC and pAUC at $p = 0.1$, approximately four points above the official Selective Mahalanobis baseline. Performance is strongest on stationary signals (fan, valve) and weakest on impulsive transients (gearbox, slider).

Index Terms— Anomalous sound detection, Mahalanobis distance, domain generalization, statistical learning, constrained computing

1. INTRODUCTION

DCASE 2026 Task 2 evaluates unsupervised anomalous sound detection (ASD) for machine condition monitoring under noisy conditions [3]. This year introduces two synchronized microphone channels recorded at different distances from the target machine, and follows the first-shot evaluation protocol introduced in [4], where the evaluation machine types are disjoint from those of the development set. Per-machine hyperparameter tuning is therefore not possible.

This submission was developed under significant practical constraints: no dedicated training infrastructure, no persistent compute environment, and only free-tier cloud notebooks with CPU. These constraints shaped the methodology. We deliberately chose a parameter-free statistical pipeline that fits inside a single session lifetime and is robust to environment interruptions, rather than a learned approach requiring extensive optimization.

2. SYSTEM DESCRIPTION

2.1. Feature extraction

For each clip we form three signals: channel 1 (near microphone), channel 2 (far microphone), and the difference (ch1 – ch2). The difference channel attenuates diffuse environmental noise while preserving energy spatially localized to the near microphone, exploiting the dual-channel structure introduced this year.

For each signal we compute two complementary time–frequency representations: a log-mel spectrogram (128 bands, 1024-point STFT, 512-sample hop, 50 Hz–8 kHz, in decibels) and a constant-Q transform (84 bins, 12 bins per octave from C1). Features are cropped or zero-padded to 313 time frames and standardized using training-set statistics.

2.2. Multi-scale temporal pooling

Each clip is summarized with multi-scale statistics. For scales $s \in \{1, 2, 4\}$, we partition the 313 frames into s equal segments and compute per-feature mean and standard deviation in each segment, yielding seven $(1 + 2 + 4)$ segment summaries with both first- and second-order moments.

2.3. L2-normalized domain-aware Mahalanobis scoring

Pooled descriptors are L2-normalized prior to distance computation. This step proved critical in our development experiments, removing amplitude-related variation that is uninformative of anomaly status. For each section, we fit a Mahalanobis model on source-domain training clips, with precision matrix estimated by Ledoit–Wolf shrinkage [5]. We compute two distances per test clip z — to the source-domain mean μ_s and the target-domain mean μ_t — and take the minimum:

$$s(z) = \min\left\{(z-\mu_s)^\top \Sigma^{-1}(z-\mu_s),\ (z-\mu_t)^\top \Sigma^{-1}(z-\mu_t)\right\}. \tag{1}$$

This implements a domain-generalization heuristic suited to the cross-domain evaluation protocol of [4].

2.4. Rank-based fusion and decision

The mel and CQT streams are fused by converting each to within-test-set ranks and averaging: $s_{\text{final}}(z) = (r_{\text{mel}}(z) + r_{\text{cqt}}(z))/2N$. Binary decisions use the per-section median as threshold.

Table 1: Development set results (%). Source/target AUC and pAUC at $p = 0.1$.

Machine	AUC src	AUC tgt	pAUC
ToyCarEmu	56.08	58.48	51.84
ToyCar	65.28	65.56	52.79
bearingEmu	63.52	57.56	57.08
fan	94.80	89.16	73.05
gearboxEmu	55.34	50.12	50.53
sliderEmu	63.96	56.16	50.97
valveEmu	80.36	71.76	54.00
Harmonic mean		**60.76**	

3. EXPERIMENTAL SETUP

We use the 1000 training clips per machine type from the development dataset [3], comprising ToyADMOS2-derived [1] and MIMII-derived [2] machines, and from the additional training dataset for the five evaluation machines. No external datasets or pre-trained models are used. All processing was performed on free-tier cloud notebooks (CPU only). The system has no learned parameters; all distance models are fit at test time from training-set statistics.

4. RESULTS

Table 1 reports per-machine performance on the seven development machine types. Across all metrics, the harmonic mean is 60.76, approximately four points above the official Selective Mahalanobis baseline. Performance is strongest on stationary signals (fan: AUC source 94.80, pAUC 73.05; valve: AUC source 80.36) and weakest on impulsive transients (gearbox: AUC source 55.34; slider: AUC source 63.96), consistent with the relative difficulty of impulsive-event regimes for distance-based scorers operating on pooled descriptors.

5. DISCUSSION

The system has clear methodological limits: it does not learn temporal context, does not exploit attribute labels where available, and does not adapt the distance metric to impulsive spectro-temporal patterns. In preliminary experiments, AudioSet-pretrained models did not outperform our hand-engineered statistics on the development data — consistent with the observation that pretraining on between-class discriminative tasks does not transfer to within-class normality estimation, the relevant structure for unsupervised cross-domain ASD.

6. REFERENCES

[1] 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 *Proc. DCASE Workshop*, 2021, pp. 1–5.

[2] 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 *Proc. DCASE2022 Workshop*, 2022.

[3] T. Nishida, N. Harada, D. Takeuchi, et al., "Description and Discussion on DCASE 2026 Challenge Task 2: Noise-Aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," *arXiv:2606.01578*, 2026.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-Shot Anomaly Detection for Machine Condition Monitoring: A Domain Generalization Baseline," in *Proc. 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.

[5] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.