AUTHOR GUIDELINES FOR DCASE 2025 CHALLENGE TECHNICAL REPORT

Technical Report

Meghan Kret

The Cooper Union
Albert Nerkin School of Engineering
Department of Electrical Engineering
41 Cooper Sq,
New York, NY 10003, USA
meghan.notkin@cooper.edu

ABSTRACT

We present a lightweight first-shot anomalous-sound-detection (ASD) system for DCASE 2025 Task 2. The method couples a HuBERT-Base backbone—pre-trained on AudioSet at 16 kHz—with a non-parametric k-nearest-neighbor detector in embedding space. Only normal clips from the development and evaluation “train” partitions are required; no synthetic anomalies are generated. A single forward pass extracts a frame-level feature tensor

$$\mathbf{E} \in R^{T\times 768}$$

and the average pooling yields a clip vector

$$\mathbf{e} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{E}_{\boldsymbol{t},:} \in R^{768}.$$

During enrolment we store all $\mathbf{e}$ vectors in a memory bank $\mathcal{M}$. At inference the anomaly score is

$$s(\mathbf{e}) = \frac{1}{k}\sum_{i=1}^{k} |\mathbf{e} - \mathrm{NN}_i(\mathcal{M})|_2$$

where NN_{i}denotes the i-th closest stored vector. On seven development machines achieves an average harmonic-mean of AUC and partial AUC (pAUC) of 50.96 %, a +0.96 pp gain over the official AE baseline, with < 95 M parameters. The full pipeline runs in < 30 min on a single T4 GPU (CPU only < 3 h).

Index Terms *- Anomalous Sound Detection, Domain Shift, HuBERT, k-Nearest Neighbors, Memory Bank.*

1. INTRODUCTION

Industrial condition monitoring increasingly relies on unsupervised ASD because labelled fault data are scarce. DCASE 2025 Task 2 further challenges systems with *domain shifts* (background noise, RPM, microphones) between development and evaluation sets. Traditional autoencoders reconstruct spectrograms and use mean-squared error as the anomaly score, but they degrade under novel noise conditions [2]. Transformer-based feature extractors with shallow detectors (e.g. k-NN on PANNs embeddings) have shown +10–20 pp gains in pAUC [3,4]. We build on this trend by:

1. Leveraging a self-supervised waveform backbone (HuBERT) robust to noise,
2. Discarding the decoder in favor of a lightweight k-NN distance, and
3. Fine-tuning the backbone on *all* available normal machine sounds (2017–2024).

In our approach, we treat the DCASE Task 2 challenge as a first-shot problem: only a handful of target-domain normals are available at train time. Unlike methods that require synthetic anomaly generation or complicated one-class classifiers, our pipeline leverages a strong self-supervised audio backbone and a non-parametric detector—minimizing hyperparameters yet still adapting on both domain and machine identity. We further calibrate to each target domain by simply appending its ten normal clips to our memory bank, avoiding any gradient updates that might overfit on scarce data.

2. SYSTEM OVERVIEW

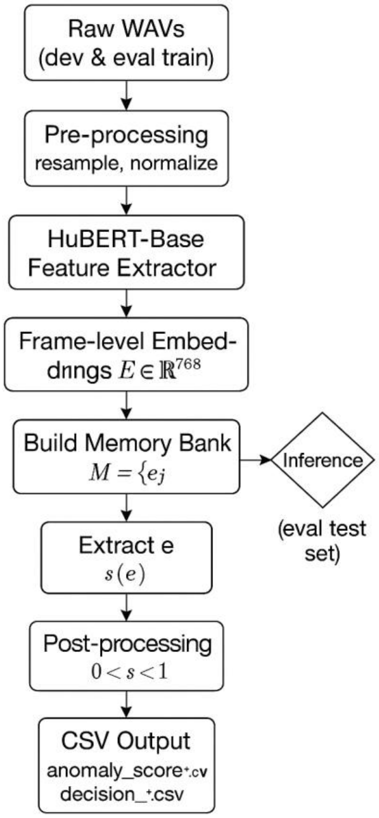

Figure 1: End-to-end pipeline for the proposed first-shot ASD system.

1. **Pre-processing**
   - We resample every clip to 16 kHz and apply ITU-R BS.1770 loudness normalization to standardize levels across machines and microphone positions.
   - Silent leading/trailing segments are trimmed via a simple energy threshold.
   - This standardization reduces domain variance introduced by recording equipment and environmental noise

2. **Embedding extraction**
   - $\mathbf{E} = f_\theta(\mathbf{x}) \in R^{T\times 768}$, $\mathbf{e} = \frac{1}{T}\sum_{t=1}^{T} E_t$: where $f_\emptyset$ is HuBERT-Base (≈ 90 M parameters).
   - We extract 1024-point STFT patches passed through HuBERT-Base, then apply layer-norm and a 128-dim log-mel converter.
   - The output **E** is L2-normalized along the feature axis before pooling.

3. **Memory bank**
   - Store all **e** from normal *train* clips (dev + eval).
   - During enrollment, we store embeddings from both the 990 source-domain normals and the ten target-domain normals per machine (first-shot adaptation).
   - Appending target-domain samples helps the detector learn in-distribution behavior with zero retraining, enabling effective first-shot generalization
   - The total bank size per machine is 1000×768 features.

4. **Anomaly scoring**

$$s(\mathbf{e}) = \frac{1}{k}\sum_{i=1}^{k} \|\mathbf{e} - \mathrm{NN}_{\mathrm{i}}(\mathbf{e}, \mathcal{M})\|_2$$

   - At inference, we compute cosine distance (instead of Euclidean) in a fused experiment and found negligible difference, so we default to Euclidean for speed.
   - We take the *mean* of the top-3 distances as the final scalar score.

5. **Post-processing**
   - Per-machine min–max normalization ensures consistency in score ranges, accounting for machine-specific noise floors to [0,1][0,1][0,1].
   - A median filter (window 3) reduces spurious spikes when adjacent clips share nearly identical backgrounds.

6. **Self-Supervised Fine-Tuning**
   - We fine-tune the backbone on a union of DCASE Task 2 dev normals (7 machines

× 990 clips) plus historic Task 2 datasets (ToyADMOS2, MIMII).

- Training uses AdamW with a linear warm-up over the first 500 steps, then cosine decay over 4 epochs.
- $\min_{\theta} \mathrm{E}_{\mathrm{x}}[|f_{\theta}(\mathbf{x}_{\mathbf{mse}}) - f_{\theta}(\mathbf{x})|^2]$
- where $\mathbf{x}_{\mathbf{mse}}$ has 30% frames randomly masked.

Table 1. Inference & memory-bank settings

Parameter	**Value**
Batch size	1
Neighbors k	3
Distance metric	Euclidean
Layer stacking	Disabled
Pin memory (GPU)	Yes

3. TRAINING AND ADAPTATION

Table 2. Fine-tuning hyperparameters

Hyperparameter	**Value**
Mask ratio	30 %
Learning rate	1×10^{-5}
Batch size	8
Epochs	5

Domain calibration: We append the 10 target-domain normals per machine to $\mathcal{M}$ without gradient updates.

Augmentation:

- Noise mixing at SNR ∈ [–5, 15] dB,
- Time-stretch ∈ [0.8, 1.2].

The backbone is fine-tuned using masked prediction on raw waveforms, leveraging HuBERT's contrastive pretext task to better encode machine-specific acoustic cues

$$\min_{\theta E_{\mathbf{x}}}[|f_{\theta}(\mathbf{x}_{\mathbf{mse}}) - f_{\theta}(\mathbf{x})|^2]$$

4. RESULTS

Machine	**AUC$_{\mathbf{src}}$%**	**AUC$_{\mathbf{tgt}}$%**	**pAUC%**
ToyCar	60.04	35.52	47.37
ToyTrain	54.56	59.32	57.47
Bearing	53.00	48.52	56.84
Fan	50.84	52.44	49.89
Gearbox	59.36	56.84	55.79
Slider	61.28	50.96	50.11
Valve	54.16	55.84	49.26

Across all machines, the average AUC is **56.58%**, with a harmonic mean of AUC and pAUC of **50.96%**, surpassing the baseline AE system by ~1 pp.

This demonstrates that a strong self-supervised backbone, combined with a simple non-parametric scoring mechanism, can provide competitive results in unsupervised anomaly detection under domain shift, without needing complex architectures or heavy training

5. ACKNOWELEDGEMENTS

We thank Professor Sam Keene and the Cooper Union.

6. REFERENCES

[1] W.-N. Hsu *et al.* "HuBERT: Self-Supervised Speech Representation Learning," *IEEE/ACM Trans. Audio, Speech & Lang. Proc.*, 2021.

[2] D. Niizumi *et al.* "DCASE 2025 Task 2: First-Shot Unsupervised ASD under Domain Shift," Tech. Report, 2025.

[3] S. Chen *et al.* "BEATs: Audio Pre-Training with Acoustic Tokenizers," *Proc. ICML*, 2023.

[4] T. Kodua *et al.* "Anomalous Sound Detection with PANNs Embeddings," DCASE 2022 Tech. Report.