CHANNEL-ROBUST EMBEDDING ENSEMBLES FOR DCASE 2026 TASK 2
Technical Report

Xing Wu

MCPX
wuuuu6767@163.com

ABSTRACT

DCASE 2026 Task 2 evaluates first-shot unsupervised anomalous sound detection for machine condition monitoring under unseen machine types and noisy two-channel recordings. This report describes four submitted systems built from three complementary normal-sound embedding models: a tiny-size FISHER industrial-signal adapter, an EAT audio-transformer adapter trained with an angular-margin objective, and a small-size FISHER industrial-signal adapter. The primary system uses only the near microphone and fuses the three model scores after fixed machine-wise normalization. The remaining systems add a conservative Wiener-enhanced channel, a single-model near-microphone fallback, and an aggressive channel-2 hedge. All systems use normal-only embedding learning and cosine one-nearest-neighbor scoring against source and target normal references; larger scores indicate sounds farther from the normal training manifold.

Index Terms— anomalous sound detection, machine condition monitoring, first-shot generalization, channel fusion, nearest-neighbor scoring

1. INTRODUCTION

Anomalous sound detection (ASD) for machine condition monitoring aims to identify abnormal operating sounds when only normal recordings are available for training. DCASE 2026 Task 2 keeps this unsupervised setting but adds a noise-aware two-channel condition and evaluates first-shot generalization to machine types that differ from the development set [1]. Its development data and evaluation philosophy build on ToyADMOS2, MIMII DG, and first-shot domain-generalization studies for machine-condition monitoring [2, 3, 4]. A practical submission must therefore be conservative about machine-specific tuning while still exploiting the far microphone when it provides useful noise information.

Our submission uses a common scoring principle across all four systems: learn normal-sound embeddings, measure the distance from each test clip to normal source and target references, and combine complementary model or channel scores only after fixed score normalization. The implementation and package generation are based on ASD-Kit [5]; this report describes our representation adaptation, anomaly scoring, and channel-fusion choices on top of that toolkit. The four package systems are Channel-1 Score Trio (XingWu_MCPX_task2_1), Wiener-Hedged Score Trio (XingWu_MCPX_task2_2), EAT ArcFace Anchor (XingWu_MCPX_task2_3), and Ch2-Hedged Score Trio (XingWu_MCPX_task2_4). They differ in how much they trust the far microphone.

2. PROPOSED METHOD

2.1. Signal notation and problem formulation

Let $x_i^{(1)} \in \mathbb{R}^L$ and $x_i^{(2)} \in \mathbb{R}^L$ denote the near-microphone and far-microphone waveforms of clip i, respectively, and let $m(i)$ be its machine type. For a representation model k, an embedding network $f_k(\cdot)$ maps a waveform to a D_k-dimensional clip vector,

$$z_{i,k}^{(c)} = f_k(x_i^{(c)}) \in \mathbb{R}^{D_k}, \qquad c \in \{1, 2\}. \tag{1}$$

The anomaly score $A(x_i)$ is a continuous value in which a larger value means the clip is farther from normal training evidence. A binary decision can be obtained with a fixed machine-dependent threshold ϕ_m:

$$\text{Decision}(x_i) = \begin{cases} \text{anomaly}, & A(x_i) > \phi_{m(i)}, \\ \text{normal}, & \text{otherwise.} \end{cases} \tag{2}$$

The submitted systems use the continuous scores for official evaluation and use thresholding only after score generation.

2.2. Normal-sound embedding learning

Each representation model is trained using only normal recordings from the task training data. Machine and section metadata define proxy classes $c_i \in \{1, \ldots, C\}$. For the FISHER-based models [6], the embedding is optimized with trainable sub-cluster AdaCos [7]. With Q subclusters per proxy class, normalized embedding $\bar{z} = z/\|z\|_2$, normalized center $\bar{w}_{c,q} = w_{c,q}/\|w_{c,q}\|_2$, and AdaCos scale α, the posterior for proxy class c is

$$P(c|x_i) = \sum_{q=1}^{Q} \frac{\exp(\alpha \bar{z}_i^\top \bar{w}_{c,q})}{\sum_{c'=1}^{C} \sum_{q'=1}^{Q} \exp(\alpha \bar{z}_i^\top \bar{w}_{c',q'})}. \tag{3}$$

The objective is the negative log likelihood,

$$\mathcal{L}_{\text{SC}} = -\frac{1}{N} \sum_{i=1}^{N} \log P(c_i|x_i). \tag{4}$$

This loss keeps normal clips of the same proxy condition compact while allowing several subclusters for different operating modes.

The EAT model [8] uses the same normal-only proxy-label training principle but with an ArcFace-style angular-margin classification objective [9]. For normalized class weight $\bar{w}_{c_i}$, normalized embedding $\bar{z}_i$, angular margin μ, and scale γ, the target logit is

$$\ell_{i,c_i} = \gamma \cos(\arccos(\bar{w}_{c_i}^\top \bar{z}_i) + \mu), \tag{5}$$

and non-target logits use $\gamma \bar{w}_c^\top \bar{z}_i$. This objective gives the single EAT model a sharper class separation than ordinary softmax training while still using no anomaly labels.

2.3. Cosine source-target nearest-neighbor scoring

After training, each model extracts normal-reference embeddings for every machine. Let $\mathcal{R}_{m,\mathrm{src}}^{(k)}$ and $\mathcal{R}_{m,\mathrm{tgt}}^{(k)}$ be the source-domain and target-domain normal reference sets for model k and machine m. Every reference and query embedding is L2-normalized before scoring. For an input waveform u, the source distance is

$$d_{\mathrm{src}}^{(k)}(u) = \min_{r \in \mathcal{R}_{m,\mathrm{src}}^{(k)}} \left\| \bar{f}_k(u) - \bar{r} \right\|_2, \tag{6}$$

and the target distance $d_{\mathrm{tgt}}^{(k)}(u)$ is computed in the same way when target-domain normal references are available. The model-level anomaly score is

$$a_k(u) = \begin{cases} \min\{d_{\mathrm{src}}^{(k)}(u), d_{\mathrm{tgt}}^{(k)}(u)\}, & |\mathcal{R}_{m,\mathrm{tgt}}^{(k)}| > 0, \\ d_{\mathrm{src}}^{(k)}(u), & \text{otherwise.} \end{cases} \tag{7}$$

Target-domain references are conservatively balanced by synthetic interpolation in the normal embedding space before nearest-neighbor scoring. This preserves the one-neighbor distance rule while reducing sensitivity to sparse target-domain normal examples.

2.4. Wiener enhancement and channel scores

For systems that use the far microphone, Ch2 is treated as a noise-reference signal rather than as a replacement for Ch1, following the Wiener-filtering principle of attenuating components dominated by estimated noise power [10]. Let $Y_i^{(1)}(t,f)$ and $Y_i^{(2)}(t,f)$ be the short-time Fourier transforms of the near and far channels. The far-channel noise power estimate is

$$\widehat{P}_{n,i}(t,f) = \eta |Y_i^{(2)}(t,f)|^2, \tag{8}$$

where η is a fixed calibration factor. A near-channel speech-and-machine power estimate is formed as

$$\widehat{P}_{s,i}(t,f) = \max\left(|Y_i^{(1)}(t,f)|^2 - \widehat{P}_{n,i}(t,f), 0 \right). \tag{9}$$

The Wiener gain is then

$$H_i(t,f) = \frac{\widehat{P}_{s,i}(t,f)}{\widehat{P}_{s,i}(t,f) + \widehat{P}_{n,i}(t,f) + \epsilon}, \tag{10}$$

and the enhanced near-channel waveform is reconstructed from

$$\widetilde{Y}_i^{(1)}(t,f) = H_i(t,f) Y_i^{(1)}(t,f). \tag{11}$$

This produces a second Ch1-derived view $\widetilde{x}_i^{(1)}$ that suppresses noise components correlated with the far microphone while preserving the near-microphone machine signal.

Table 1: Submitted system definitions.

System name	System definition	Fusion rule
Channel-1 Score Trio	Ch1 three-model ensemble using FISHER-tiny, EAT ArcFace, and FISHER-small	mean of fixed-normalized raw Ch1 scores
Wiener-Hedged Score Trio	Ch1 plus Wiener ensemble using the same three models	mostly raw Ch1 with 0.25 Wiener weight
EAT ArcFace Anchor	single-model Ch1 fallback using the EAT ArcFace model	no model fusion
Ch2-Hedged Score Trio	aggressive Ch2/Wiener hedge using the same three-model base	0.60 raw Ch1, 0.30 Wiener Ch1, 0.10 direct Ch2

2.5. Fixed score normalization and fusion

Scores from different models and channels are fused after a fixed machine-wise normalization. For each machine m, model k, and channel view v, a calibration set of normal scores $\mathcal{C}_{m,k,v}$ is formed before evaluating the submitted test clips. A raw score $a_k(x_i^{(v)})$ is converted to the fixed normalized score

$$\widetilde{a}_{k,v}(x_i) = \frac{1 + \sum_{c \in \mathcal{C}_{m(i),k,v}} \mathbb{I}\{c \le a_k(x_i^{(v)})\}}{|\mathcal{C}_{m(i),k,v}| + 1}. \tag{12}$$

This transform is fixed by the calibration scores and is not recomputed from the submitted test set. It keeps larger values aligned with higher abnormality while reducing scale differences between EAT and FISHER embeddings.

Let $\mathcal{K} = \{\text{FISHER-tiny}, \text{EAT-ArcFace}, \text{FISHER-small}\}$ denote the three submitted embedding models. The raw Ch1 ensemble score is

$$A_{\mathrm{raw}}(x_i) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \widetilde{a}_{k,\mathrm{ch1}}(x_i). \tag{13}$$

The Wiener-enhanced Ch1 ensemble score $A_{\mathrm{wiener}}(x_i)$ is computed by replacing $x_i^{(1)}$ with $\widetilde{x}_i^{(1)}$, and the direct far-channel ensemble score $A_{\mathrm{ch2}}(x_i)$ is computed from $x_i^{(2)}$.

2.6. Submitted systems

Channel-1 Score Trio. The main system uses only the near microphone. Its score is

$$A_1(x_i) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \widetilde{a}_{k,\mathrm{ch1}}(x_i). \tag{14}$$

This system is the most conservative submission because Ch1 is the closest microphone to the machine and is real for both ordinary and emulated machine conditions. The three-model normalized average combines an industrial tiny FISHER representation, a general-audio EAT representation, and an industrial small FISHER representation without relying on any Ch2 assumption.

Wiener-Hedged Score Trio. The second system keeps the same three-model raw Ch1 ensemble but adds a small Wiener-enhanced term:

$$A_2(x_i) = 0.75A_{\text{raw}}(x_i) + 0.25A_{\text{wiener}}(x_i). \tag{15}$$

This system tests the hypothesis that Ch2 is most useful as a noise reference. The small Wiener weight lets the enhanced signal help when far-channel noise estimation is reliable while keeping the raw near-microphone evidence dominant.

EAT ArcFace Anchor. The third system is an independent simple fallback:

$$A_3(x_i) = \widetilde{a}_{\text{EAT-ArcFace,ch1}}(x_i). \tag{16}$$

It uses only Ch1 and only the EAT angular-margin model. Its purpose is to provide a clean non-ensemble submission whose behavior is easier to inspect if the multi-model systems overfit the development distribution.

Ch2-Hedged Score Trio. The fourth system uses the same three-model base but assigns more mass to the noise-aware views:

$$\begin{aligned} A_4(x_i) = {} & 0.60A_{\text{raw}}(x_i) + 0.30A_{\text{wiener}}(x_i) \\ & + 0.10A_{\text{ch2}}(x_i). \end{aligned} \tag{17}$$

This system is intentionally the most channel-adaptive member of the slate. It preserves the raw Ch1 majority vote, gives substantial weight to Wiener-enhanced Ch1, and includes a small direct Ch2 score to capture possible ONSTAGE cases where far-field information carries discriminative machine-condition cues.

2.7. Development-set metrics

Table 2 reports the selected development-test summary from the signed-log robust-z four-metric candidate table. All values are percentages, and each summary value is the harmonic mean across the seven labeled development machines. The results show that the two channel-aware score trios keep similar overall performance to the Ch1-only trio while improving the pAUC summary slightly; the single EAT anchor remains a simpler fallback rather than the strongest development-set scorer.

Table 2: Development-test summary for the four submitted systems.

Metric	Ch1 Trio	Wiener Trio	EAT Anchor	Ch2 Trio
AUC_s	65.36	67.14	60.04	67.17
AUC_t	61.25	61.84	57.66	61.24
pAUC	53.43	53.76	52.06	53.77
hmean	59.59	60.40	56.38	60.23

Table 3 gives the corresponding per-machine hmean values.

Table 3: Per-machine development-test hmean values.

Machine	Ch1 Trio	Wiener Trio	EAT Anchor	Ch2 Trio
bearingEmu	57.22	57.42	56.53	57.28
fan	56.69	57.73	47.75	57.53
gearboxEmu	64.34	65.29	63.73	64.77
sliderEmu	52.56	54.01	51.15	53.42
ToyCar	70.33	70.28	60.19	69.14
ToyCarEmu	60.45	60.58	59.29	61.40
valveEmu	58.74	60.30	59.54	60.67

The Wiener-Hedged Score Trio gives the best across-machine hmean, while the Ch2-Hedged Score Trio is retained because it has the strongest ToyCarEmu and valveEmu hmean among the four systems and provides the most explicit hedge for far-channel benefit.

2.8. Representation models

The EAT member adapts a base-sized audio transformer [8] using the CLS token as the clip representation and a linear projection to the anomaly-detection embedding. Low-rank adapters [11] are inserted into the attention projections so that the task adaptation changes the representation without fully overwriting the base audio-transformer structure.

The two FISHER members adapt industrial-signal foundation models [6]. Each model computes log-magnitude STFT features internally, normalizes them with fixed FISHER constants, reshapes the output into sub-band embeddings, and pools the sequence by an attention-statistic projection before normal-only discriminative training. The tiny and small model sizes are both retained because they provide related but not identical industrial-signal representation biases.

3. CONCLUSION

This report describes four final DCASE 2026 Task 2 systems organized around channel robustness. Channel-1 Score Trio is the primary Ch1-only three-model ensemble, Wiener-Hedged Score Trio adds a conservative Wiener-enhanced Ch1 contribution, EAT ArcFace Anchor is a single-model Ch1 fallback, and Ch2-Hedged Score Trio is an aggressive Ch2/Wiener hedge. On the selected development-test summary, the three score-trio systems obtain hmean values of 59.59, 60.40, and 60.23, respectively. The common foundation is normal-only embedding learning and cosine one-nearest-neighbor scoring, while the submitted diversity comes from model family, channel usage, and fixed score-level fusion.

4. REFERENCES

[1] 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," *arXiv e-prints: 2606.01578*, 2026.

[2] 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. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021, pp. 1–5.

[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 *Proc. 7th Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2022.

[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] T. Fujimura, K. Wilkinghoff, K. Imoto, and T. Toda, "ASDKit: A toolkit for comprehensive evaluation of anomalous sound detection methods," in *Proc. 10th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Barcelona, Spain, 2025, pp. 40–44, doi: 10.5281/zenodo.17251589.

[6] P. Fan, A. Jiang, S. Zhang, Z. Lv, B. Han, X. Zheng, W. Liang, J. Li, W.-Q. Zhang, Y. Qian, X. Chen, C. Lu, and J. Liu, "FISHER: A foundation model for multi-modal industrial signal comprehensive representation," *arXiv preprint arXiv:2507.16696*, 2025.

[7] K. Wilkinghoff, "Combining multiple distributions based on sub-cluster AdaCos for anomalous sound detection under domain shifted conditions," in *Proc. DCASE Workshop*, 2021.

[8] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 3807–3815, doi: 10.24963/ijcai.2024/421.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[10] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Cambridge, MA, USA: MIT Press, 1949, doi: 10.7551/mitpress/2946.001.0001.

[11] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.