DOMAIN-BLIND, DUAL-MICROPHONE ENSEMBLES FOR FIRST-SHOT ANOMALOUS SOUND DETECTION

Technical Report

Junhyeong Kwon, Jongsuk Choi

Korea Institute of Science and Technology, Seoul, Korea
meoruu00@gmail.com, cjs@kist.re.kr

ABSTRACT

We describe our submission to DCASE 2026 Task 2 (Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring), the first edition to release two-channel near- and far-microphone recordings, in the first-shot setting under domain shift. With only about a thousand normal clips per machine and no anomalies, we keep the analytical backbone *frozen*: a self-supervised BEATs encoder used as a fixed feature extractor, avoiding the representation collapse that full fine-tuning invites at this data scale. The core anomaly score is a per-domain-centred Mahalanobis ensemble over its embeddings. We exploit the second microphone in two unsupervised ways: a magnitude spectral-subtraction front-end that suppresses shared ambient noise, and a cross-channel difference stream from a second frozen encoder, Dasheng. Because each test clip's domain is hidden at scoring time, we score every clip against *both* the source and target normal models and take the minimum—a domain-blind rule that is a correctness condition on the tagless evaluation set. Finally, as a separate decorrelated member we add a parameter-efficient LoRA–ArcFace encoder scored by cosine k-NN, our largest single *learned* gain. On the development set the full ensemble reaches an official Ω of **68.97 %**, a $+19\,\%$ relative gain over the official Mahalanobis baseline; even our training-free core reaches $67.85\,\%$. We submit four systems spanning a training-free baseline through the full learned ensemble.

***Index Terms*—** anomalous sound detection, first-shot, domain-blind scoring, frozen encoder, Mahalanobis ensemble, dual-microphone, LoRA–ArcFace

1. INTRODUCTION

DCASE 2026 Task 2 [1] asks for unsupervised anomalous sound detection (ASD) in the *first-shot* setting [2]: only *normal* training clips are available, no anomalies, and the test set mixes a large *source* domain of roughly 990 normal clips per machine with a sparse *target* domain of about 10 clips, under unknown acoustic noise. The task data derive from the ToyADMOS2 [3] and MIMII DG [4] corpora. The 2026 edition is the first to provide two synchronised channels, a near and a far microphone, opening a new and largely unexplored axis for noise-robust detection.

First-shot ASD systems cluster into two families. *Inlier-modeling* methods score a clip by its distance or likelihood under a model of the normal data; the official autoencoder and Mahalanobis baselines [1] are of this kind. *Discriminative* methods instead learn an embedding by classifying machine and attribute metadata, often with a sub-cluster ArcFace or AdaCos objective [5], and score by distance in that space. A recent trend keeps a large self-supervised encoder *frozen* and reuses its embeddings with a simple distance backend, which is robust at the few-hundred-clip scale of this task. Our design combines all three: a frozen-encoder inlier-modeling core, a small discriminative LoRA–ArcFace member, and the second microphone with a domain-blind scoring rule layered on top.

Two facts shape our design. First, the official ranking metric Ω is the *global harmonic mean* over the pooled set of per-machine $\{\mathrm{AUC}_{\mathrm{source}}, \mathrm{AUC}_{\mathrm{target}}, \mathrm{pAUC}\}$ values; it is dominated by the *worst* cell, so robustness across machines matters more than peak performance on any one. Second, AUC and pAUC are rank-based, so any per-machine monotonic rescaling of scores is inert—only the within-machine ordering and cross-machine variance can be improved.

Our contributions are:

- a competitive *training-free* frozen-encoder Mahalanobis ensemble for the two-channel first-shot setting;
- two complementary ways of *exploiting the second microphone*—a magnitude-subtraction front-end and a cross-channel difference embedding;
- a *domain-blind* source/target-minimum scoring rule that stays correct when the test-clip domain is unobserved; and
- a parameter-efficient *LoRA–ArcFace member* used as a decorrelated ensemble increment rather than a standalone detector.

2. SYSTEM DESCRIPTION

All audio is resampled to 16 kHz. The near microphone ch1 is the primary signal; the far microphone ch2 both exposes the shared ambient field for the spectral-subtraction front-end of Sec. 2.2 and, in the analytical core, serves as a complementary discriminative stream through a joint Mahalanobis on the concatenated ch1 and ch2 embeddings, cross-channel cluster geometry, and cross-channel level features. Embeddings are mean-pooled over time, since the official metric is computed at clip level. Figure 1 summarises the four submitted systems.

The analytical core is built on two *frozen* encoders. **BEATs** [6], an AudioSet-pre-trained 90.35 M-parameter iter3+ model, supplies the final-layer mean-pooled 768-d embedding of ch1, with intermediate layers feeding specific sub-scores. **Dasheng** [7], an 85.45 M-parameter masked-autoencoder, is used in one submitted system on the cross-channel difference. No backbone weights are updated for the core: with only $\sim$1000 normal clips per machine and no anomalies, *full* fine-tuning of a 90 M-parameter transformer is prone to representation collapse and source-domain overfitting, so only the ENS member of Sec. 2.4 is trained.

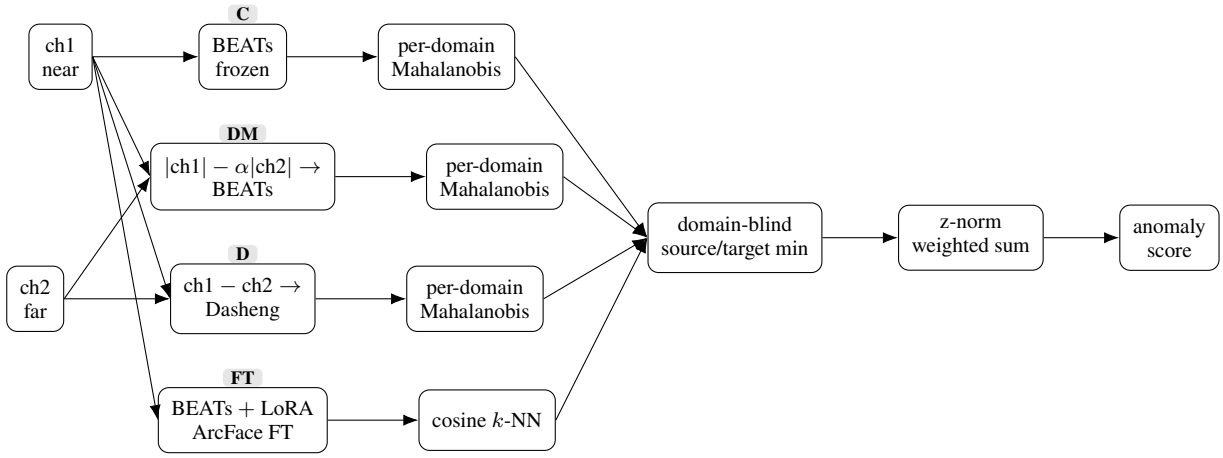

Figure 1: Overview of the four submitted systems. One frozen BEATs backbone is reused across the Mahalanobis core, the dual-microphone front-end, and the LoRA–ArcFace member; every stream is scored with the domain-blind source/target minimum before z-normalised weighted-sum fusion.

2.1. ULICP: per-domain Mahalanobis ensemble

The analytical core scores a test embedding by its Mahalanobis distance to the normal training manifold, with two design choices adapted to domain shift. *Per-domain centring:* source and target training clips are centred by their own means before a single shared covariance basis is fitted by weighted PCA, with target clips up-weighted to compensate for the $\sim$100:1 imbalance, so the basis represents both domains while keeping the sparse target influential. *Top-k PCA Mahalanobis:* a diagonal Mahalanobis in the leading principal subspace for numerical stability in 768 dimensions.

On top of this base we fuse, by z-normalised weighted sum, a small family of complementary sub-scores:

- a 4-cluster boundary-geometry score;
- a change-point anomaly proxy on the per-channel score sequences;
- a joint cross-channel Mahalanobis on the concatenated ch1 and ch2 embeddings;
- a cross-channel residual-conformity score from a ridge ch2$\rightarrow$ch1 prediction residual, ranked per domain;
- a local Mahalanobis density, computed un-centred and therefore domain-invariant, used as a log-ratio denominator.

The fusion weights are small and were fixed on the development set; the ensemble is deliberately conservative to avoid over-fitting the seven development machines.

2.2. ULICPDS and ULICPD: exploiting the second microphone

We use the far microphone in two orthogonal ways. *Dual-microphone front-end, ULICPDS.* Anomalies are near-field events on ch1, while much of the noise is shared between the microphones, so we form a magnitude-subtracted signal $|\text{ch1}| - \alpha\,|\text{ch2}|$ in the spectral-magnitude domain and embed it through the same frozen BEATs at an intermediate layer; a per-domain Mahalanobis on this "cleaned" stream is added with a small weight. This front-end is unsupervised—no labels, no training, only the two channels. *Cross-channel difference, ULICPD.* Separately, we embed the cross-channel difference ch1 $-$ ch2 with the frozen **Dasheng** encoder and add a per-domain Mahalanobis on it. Being a different encoder on a different signal, this stream is decorrelated from the BEATs streams and provides a small, diversity-driven gain.

2.3. Domain-blind source/target-minimum scoring

The development filenames reveal each test clip's domain, but on the **evaluation set the domain is hidden**. A scorer that routes by the missing domain tag would, by default, centre every evaluation clip on the sparse $\sim$10-clip target mean and apply the target-domain fusion weights, mis-centring the $\sim$50 % of clips that are actually source-domain: a normal source clip, displaced by the source$\rightarrow$target mean gap, is pushed to a larger Mahalanobis distance and looks anomalous. The shared covariance basis, fitted on both domains, is unaffected; the degeneracy lies in the per-clip centring and fusion weighting. We therefore score *every* clip against *both* the source-centred and the target-centred model and z-normalise each into $s_{\text{src}}(x)$ and $s_{\text{tgt}}(x)$. The domain-blind score is their element-wise **minimum**, the distance to the nearest normal manifold:

$$s(x) = \min\big(s_{\text{src}}(x),\ s_{\text{tgt}}(x)\big). \tag{1}$$

This rule uses no test-time domain knowledge and is applied identically on development and evaluation. All dominant streams use it—the per-domain Mahalanobis core, the dual-microphone front-end, the Dasheng cross-channel difference, and the fine-tuned cosine k-NN—while the un-centred local density is already domain-invariant, so the minimum is a no-op for it. Two sub-scores are exceptions: the cross-channel residual-conformity score, weight 0.20 within the analytical core, keeps an all-target default on the tagless evaluation set, and a small frame-level change-point term of weight 0.05 is omitted from the redeployed core. Because the dominant streams are scored domain-blind, the composite is *predominantly*, though not fully, domain-blind, and the development gain we report for this rule is a conservative lower bound.

2.4. ENS: LoRA–ArcFace fine-tuned member

As a decorrelated learned increment we fine-tune BEATs with **LoRA** [8], using rank-64 adapters on the query/value projections

Table 1: Submitted systems and their components.

#	Label	C	DM	D	FT	Total	Train.
1	ENS	✓	✓		✓	92.95 M	2.60 M
2	ULICPDS	✓	✓			90.35 M	0
3	ULICP	✓				90.35 M	0
4	ULICPD	✓		✓		175.80 M	0

Table 2: Development-set Ω (%); Δ = gain over baseline.

System	Ω (%)	Δ
MSE baseline [1]	57.21	-0.78
Mahalanobis baseline [1]	57.99	—
ULICP	67.85	$+9.86$
ULICPD	67.97	$+9.98$
ULICPDS	68.07	$+10.08$
ENS	**68.97**	$+$**10.98**

at scale 1.0 for 2.36 M trainable parameters, together with a metric-learning head in the spirit of sub-cluster discriminative ASD [5]. The head is an MLP $768 \rightarrow 256 \rightarrow 128$ that adds 0.24 M parameters, so ENS totals $90.35 + 2.36 + 0.24 \approx 92.95$ M; it is trained with **ArcFace** [9] at scale 30 and margin 0.4 to classify the 45 *machine* $\times$ *attribute* $\times$ *domain* combinations of the development set, with embedding mixup at $\alpha = 0.2$ for regularisation, over 25 epochs of AdamW at learning rate 10^{-3}. At test time the backbone is a feature extractor scored by cosine k-NN against the source and target normal banks, under the same domain-blind minimum. The fine-tuned score is z-normalised and added with weight 0.1. It is a *member*, not a standalone system: alone it does not beat the frozen core, but it is decorrelated enough to reduce the ensemble's per-clip ranking errors, particularly on the more mechanically separable machines.

3. EXPERIMENTS

We submit four systems spanning the risk spectrum, listed in Table 1; the team rank is taken from the best single system. They form a factorial ablation over four components: C, the BEATs per-domain Mahalanobis core; DM, the dual-microphone front-end; D, the Dasheng cross-channel stream; and FT, the LoRA–ArcFace member. Systems 2–4 use *no task training*, label-free and frozen-encoder only, while system 1 adds the LoRA–ArcFace member; all four use domain-blind scoring. ENS is a score-level ensemble that reuses *one* frozen BEATs backbone, so only the 2.60 M LoRA+MLP is added over ULICP rather than a second encoder; ULICPD alone carries a second backbone, Dasheng, hence its larger parameter total. The Train. column of Table 1 counts task-trained parameters, of which the frozen core has none.

Table 2 reports development results under the official metric, the global harmonic mean Ω over all per-machine $\text{AUC}_{\text{source}}$, $\text{AUC}_{\text{target}}$, and pAUC values in percent. Because the principal streams do not consult the development domain tags, these domain-blind source/target-minimum scores are a closer proxy for blind-evaluation behaviour than an oracle-domain figure; they remain a seven-machine development proxy, not a guarantee on the five blind machines.

All four systems clear the official Mahalanobis baseline by nearly 10 points, and the four entries are deliberately engineered as a factorial ablation in which each adjacent pair isolates one component. The training-free core ULICP already reaches $+9.86$ over baseline. Adding the unsupervised dual-microphone front-end isolates that factor, ULICPDS$-$ULICP $= +0.22$; replacing it with the Dasheng cross-channel-difference stream isolates a cross-encoder-diversity factor, ULICPD$-$ULICP $= +0.12$; and stacking the decorrelated LoRA–ArcFace member on the dual-microphone system isolates the learned-member factor, the largest single increment at ENS$-$ULICPDS $= +0.90$, for a total $+1.12$ over the training-free core. Separately, the domain-blind minimum rule is a *correctness* condition rather than a peak lever. In a development tag-strip simulation, with the development domain tags hidden to mimic evaluation, it is worth $+0.0097\,\Omega$ over the naïve all-target routing, with a 95 % bootstrap CI of $[+0.0012, +0.0188]$ over development clips and $P(\Delta > 0) = 0.986$—a conservative lower bound that mainly prevents silent degradation of source-domain clips on the blind set. To our knowledge no competing entry reports a confidence interval on a scoring-rule change; we do so because the worst-cell-dominated Ω makes single point estimates fragile.

Table 3: Per-machine development results; best per cell in bold.

	Mahalanobis baseline			ENS		
Machine	AUC_{src}	AUC_{tgt}	pAUC	AUC_{src}	AUC_{tgt}	pAUC
ToyCar	62.72	70.92	57.16	**72.32**	**89.04**	**63.95**
ToyCarEmu	57.04	72.60	**55.68**	**58.40**	**91.80**	52.47
bearingEmu	68.64	61.12	60.63	**69.20**	**67.60**	**61.32**
fan	63.48	44.20	52.26	**71.36**	**68.04**	**61.42**
gearboxEmu	66.84	57.04	52.79	**72.32**	**71.88**	**61.11**
sliderEmu	61.48	52.16	50.58	**62.44**	**71.92**	**56.74**
valveEmu	65.60	55.00	49.21	**94.80**	**89.12**	**78.42**

Table 3 gives the per-machine breakdown for ENS against the Mahalanobis baseline. ENS improves 20 of the 21 cells; the sole regression is the ToyCarEmu pAUC. The harmonic Ω is set by its weakest cells—here the ToyCarEmu source-AUC of 58.4 % and pAUC of 52.5 % on a low-coherence, moving source whose target-AUC is in fact high at 91.8 %—which is why our design prioritises lifting the worst cell over peak performance elsewhere. All numbers come from a single fixed configuration applied uniformly to every machine and domain, with no per-machine selection of scores or weights, which would inflate the figures.

4. CONCLUSION

Our full ensemble reaches $\Omega = 68.97$ % on the development set, a $+19$ % relative gain over the official Mahalanobis baseline, and the results support a simple message for first-shot two-channel ASD. A *frozen* encoder with a carefully constructed per-domain Mahalanobis ensemble is already a strong, stable baseline that needs no task training; a small, decorrelated *LoRA–ArcFace member* is the largest single learned increment, working alongside the frozen core rather than replacing it; and the *second microphone* adds small but consistent gains. Orthogonally to peak performance, the *domain-blind minimum* rule is the single most important *correctness* condition for the evaluation set, where the test-clip domain is unobserved. We did not find a reliable way to close the remaining gap to a fully label-supervised oracle without anomaly labels, consistent with the unsupervised nature of the task.

5. 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," *In arXiv e-prints: 2606.01578*, 2026.

[2] 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.

[3] 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)*, 2021, pp. 1–5.

[4] 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)*, 2022.

[5] K. Wilkinghoff, "Sub-cluster AdaCos: Learning representations for anomalous sound detection," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2023.

[6] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[7] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Proceedings of Interspeech*, 2024.

[8] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[9] 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 (CVPR)*, 2019.