Noise-Aware Reference Denoising for First-Shot Anomalous Sound Detection

Technical Report

Nam Kyun Kim

Automotive Electronics R&D Center,
Korea Automotive Technology Institute (KATECH),
Gwangju, Republic of Korea
kimnk@katech.re.kr

ABSTRACT

This report addresses the noise-aware first-shot unsupervised anomalous sound detection (ASD) task of the DCASE 2026 Challenge Task 2, where each clip has a near and a far microphone but the official baseline scores only the near channel. The far microphone is exploited as a noise reference: a per-machine noise-transfer function is built from each channel's minimum-statistics noise floor rather than the full far spectrum, so that machine sound leaking into the far channel is preserved. The near channel is then denoised by floored spectral subtraction before a reconstruction autoencoder with a Mahalanobis score, and a per-band adaptive variant that over-subtracts where the local SNR is low performs best. Every statistic is computed from training-normal clips only, and all detectors are reported as the mean±std over ten autoencoder seeds. The noise-aware denoising robustly lifts the official development score from the baseline 0.5766 to a ten-seed ensemble of 0.6470, and three train-only post-hoc re-scorings targeting the metric bottleneck reach 0.6544. In contrast, an outlier-exposure component, architecture ablations, and from-scratch fusion models do not survive seed variance and are reported as negative results. The pipeline transfers unchanged to the all-real evaluation machines.

***Index Terms*—** First-shot ASD, noise-aware, reference denoising, spectral subtraction, seed robustness.

1. INTRODUCTION

Unsupervised anomalous sound detection (ASD) aims to detect mechanical faults from sound when only normal operating sound is available for training, a setting central to condition monitoring where faults are rare and diverse. The DCASE Task 2 series has driven this problem from the early unsupervised formulation [1] through domain shift [2, 3] to the *first-shot* setting [4, 5, 6], in which the evaluation machine types are entirely novel and per-machine threshold tuning is infeasible, so a system must generalise across machines. Two families address it: reconstruction or one-class autoencoders, exemplified by the official selective-Mahalanobis baseline [4, 7]; and self-supervised or classification-based detectors with auxiliary objectives such as spectral-temporal classification [8] or outlier exposure [9] over frozen encoders such as M2D [10] or BEATs [11].

The new element of DCASE 2026 is *noise awareness*: every clip is recorded by two microphones, one near and one far from the machine [6]. The far microphone carries relatively stronger environmental noise and weaker machine sound, so it is a natural reference for the noise contaminating the near channel. Yet the official baseline discards the second microphone and scores the near channel alone, leaving the defining cue of the task unused. Used as a denoising reference, the far channel connects the task to classical speech enhancement, such as spectral subtraction [12], the MMSE log-spectral amplitude estimator [13], and minimum-statistics noise estimation [14]. To date, however, these tools have rarely been made fully *train-only* for first-shot ASD, as the concealed-label protocol requires.

This report proposes a noise-aware front-end that turns the far microphone into a train-only denoiser for a reconstruction autoencoder with a Mahalanobis score. A per-machine noise-transfer function is estimated from each channel's minimum-statistics noise floor, and the near channel is denoised by floored spectral subtraction, with an adaptive per-band variant performing best. Every detector is evaluated under a seed-robust protocol that separates genuine gains from seed luck, and a train-only post-hoc re-scoring targets the pAUC/target bottleneck without retraining. Under this protocol, the noise-aware denoising is the only added modeling component that robustly improves the official score.

2. DATASET AND TASK

DCASE 2026 Task 2 provides seven development machine types and five evaluation machine types that are mutually disjoint [6]. The seven development types follow the ToyADMOS2 [15] (ToyCar) and MIMII DG [16] (fan, gearbox,

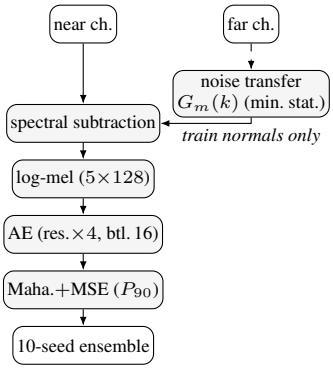

Figure 1: Noise-aware pipeline: far-channel denoising of the near channel (train-only) before a Mahalanobis-scored autoencoder (ten-seed ensemble).

bearing, slide rail, valve) lineages, provided as two-channel recordings; five are IR-simulated (*Emu*), with only fan and ToyCar non-simulated (Table 4). Each machine type supplies 990 source-domain and 10 target-domain normal training clips; the source domain dominates, so the target domain is severely under-represented. Every clip is a two-channel recording, one microphone near and one far from the machine. For the evaluation machines the source/target domain and the normal/anomaly label of each test clip are concealed, so a detector may use only the clip itself and the training data.

Performance is the official score Ω_{dev}, the harmonic mean over the seven development machines of three per-machine quantities: the source-domain AUC, the target-domain AUC, and the partial AUC at a maximum false-positive rate of 0.1. The harmonic mean penalises any weak machine, rewarding uniformly good detectors.

3. METHOD

Every component is computed independently for each machine type and each test clip, using training-normal data only; the full pipeline is shown in Fig. 1.

3.1. Reconstruction autoencoder backbone

From each channel a log-mel spectrogram (128 mel bins, a 1024-point FFT, 512-sample hop) is computed, and the autoencoder input is formed by stacking 5 consecutive frames into a 640-dimensional vector; the vectors are globally standardised using training-normal statistics only. The far channel is used solely to estimate the noise reference of Section 3.2; the autoencoder itself ingests a single (denoised) channel, matching the baseline's input dimension. The backbone (AE) is an improved variant of the DCASE 2023 dense autoencoder [4, 7], trained from scratch on the DCASE 2026 training-normal clips of each machine. The encoder compresses the 640-dimensional input to a 16-dimensional bottleneck through one projection and four residual blocks of width 512 (GELU activations, batch normalisation), and the decoder mirrors it. The clip score blends a Mahalanobis term on the per-frame reconstruction error with the reconstruction energy, aggregated over frames by the 90th percentile:

$$A(x) = (1-\alpha)\, P_{90}\big\{(e_i-\bar{e})^{\top}\Sigma^{-1}(e_i-\bar{e})\big\} + \alpha\, P_{90}\big\{\|e_i\|^2\big\}. \tag{1}$$

Here e_i is the reconstruction error of frame i, $\bar{e}$ and Σ are the mean and ridge-regularised full covariance of the training-normal errors of that machine [17], $P_{90}\{\cdot\}$ is the 90th percentile over the frames of a clip, and $\alpha{=}0.2$ weights the energy term. The clip score is standardised by a gamma fit to the training-normal scores. A Ledoit–Wolf shrinkage covariance was tried in place of Σ and adds nothing ($+0.0003\,\Omega_{\mathrm{dev}}$).

3.2. Noise-transfer reference denoising

The far microphone is a noise reference but is itself contaminated by machine sound, so subtracting its full spectrum would remove machine signal. The per-machine *noise transfer* is instead estimated by minimum statistics [14]: let $\mathrm{floor}_x(k)$ be the mean over training-normal clips of the low time-percentile (P_{20}) of the power spectrum of channel $x \in \{n, f\}$ at frequency bin k, taken where the machine is quiet; then

$$G_m(k) = \mathrm{floor}_n(k) \,/\, \mathrm{floor}_f(k) \tag{2}$$

is the near/far environmental-noise coupling of machine m, free of (loud) machine sound. Given the near and far clip power spectra $P_n(k,t), P_f(k,t)$, floored power spectral subtraction [12] yields the denoised near power

$$|\hat{S}(k,t)|^2 = \max\big(P_n - \alpha_s\, G_m(k)\, P_f,\ \beta\, P_n\big), \tag{3}$$

where the over-subtraction factor α_s and the spectral floor $\beta{=}0.10$ [18] cap machine-sound loss and suppress musical noise; the fixed version ($\alpha_s{=}1.5$) is RefSub. The adaptive AdaSub replaces the constant α_s by a per-band, per-frame value driven by the local a-posteriori SNR $\gamma_{\mathrm{dB}}(k,t){=}10\log_{10}(P_n/(G_m P_f))$, $\alpha(k,t) = \mathrm{clip}(a_0{-}0.1\,\gamma_{\mathrm{dB}}(k,t), a_{\min}, a_{\max})$, with $(a_0, a_{\min}, a_{\max}) = (3, 1, 4)$, so noise-dominated bands (low γ) are over-subtracted more and machine-dominated bands less.

3.3. Advanced front-ends

A second route uses the near–far coherence rather than the noise transfer. CohGate is an inter-channel coherence gate [20]

$$H(k,t) = \mathrm{clip}\big(1 - |\Gamma(k,t)|^2,\ \beta,\ 1\big), \tag{4}$$

where $|\Gamma|^2{=}|S_{nf}|^2/(S_{nn}S_{ff})$ is the recursively smoothed magnitude-squared coherence between channels (coherent energy $\approx$ shared environmental noise), and Combo is the

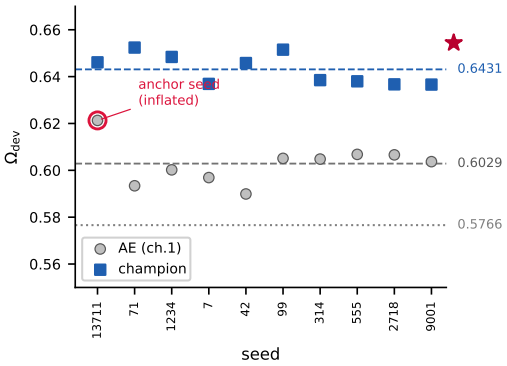

Figure 2: Per-seed Ω_{dev} for the improved AE (ch. 1, grey) and the champion (blue, submitted S1); dashed lines mark the ten-seed means, dotted the baseline. The AE anchor seed is inflated; the champion is higher and tighter, and its ten-seed ensemble ($\star$, 0.6544) exceeds every single seed.

per-clip mean of the standardised AdaSub and CohGate scores. More aggressive multiplicative gains (decision-directed Wiener [19], MMSE log-spectral-amplitude [13]) and a coherence-gated subtraction (CohSub) were also tried; all over-suppress and are not adopted. All statistics (G_m, the floor, the gamma calibration, the coherence smoothing) are train-only, so each denoised clip is a per-clip function of that clip and the training data.

4. RESULTS

4.1. Experimental setup

Each per-machine autoencoder is trained with Adam (learning rate 10^{-3}, weight decay 10^{-4}, batch 256) for 100 epochs. Single trained autoencoders are noisy: the improved AE's anchor seed reads Ω_{dev}=0.6213 against a ten-seed mean of 0.6029 (Fig. 2), and the apparent single-seed gains of a build-up ablation all fall within $\pm$0.022. Every detector is therefore trained at ten random seeds and reported as the per-seed mean$\pm$std and as a *score ensemble* that averages the ten standardised clip scores, with single-seed deltas below 0.022 treated as run-to-run variance and the configuration as a whole the only robust architectural gain ($+0.0263\pm0.0083$ over the baseline). Global hyperparameters $(\alpha_s, \beta, a_0, P_{20}, \alpha)$ are selected on the development machines and then frozen for evaluation. Because the evaluation set is entirely real, selection is anchored to the two non-simulated machines (ToyCar and fan; Table 4) and to the all-dev Ω_{dev}, rather than to any single machine; no evaluation clip is ever used. As the development and evaluation machine types are disjoint, this is model selection, not test-set tuning. Each per-clip test score passes a guard that raises on any cross-clip reduction (mean, std, percentile), so no submitted score can depend on a test-set statistic, and re-deriving the submission scores through the guard reproduces them byte-for-byte. All scores below use the official Ω_{dev}; the single-channel baseline scores 0.5666 (MSE) and 0.5766 (selective Mahalanobis) [6].

Table 1: Ten-seed development ladder, official Ω_{dev} over all seven machines: per-seed mean$\pm$std and score ensemble. No outlier exposure.

System	mean$\pm$std	ensemble
Baseline MAHALA [6]	0.5766	
AE (ch1, improved)	.6029$\pm$.0083	.5970
+RefSub	.6265$\pm$.0128	.6153
+AdaSub	.6370$\pm$.0062	.6470
+DM + shrink + P_{95}	.6431$\pm$.0057	**.6544**

Table 2: Candidate front-ends across the denoising-aggressiveness axis (ten-seed Ω_{dev}, per-seed mean$\pm$std); the fan-only column shows the coherence bet's real-machine advantage. S1/S2 submit AdaSub/RefSub, S3 the AdaSub+CohGate blend; S4 is a dev-selected ensemble (Table 3), not a single front-end.

Front-end	all-dev Ω_{dev}	fan-only Ω_{dev}
AE (ch1, improved)	.6029$\pm$.0083	.5413$\pm$.0187
RefSub	.6265$\pm$.0128	.5869$\pm$.0230
AdaSub	**.6370$\pm$.0062**	.6008$\pm$.0031
CohGate	.6035$\pm$.0044	**.6551$\pm$.0077**
Combo	.6182$\pm$.0100	.6519$\pm$.0114

4.2. Results and discussion

The noise-aware front-end is the robust gain. Table 1 reports the ten-seed ladder. Using the far channel as a train-only noise reference improves the autoencoder monotonically and robustly: the adaptive AdaSub reaches Ω_{dev}=0.6470, $+0.0704$ over the baseline, and its score ensemble exceeds its per-seed mean (so the ensemble is the right system to submit). Table 2's fan-only column shows the same gain on fan, the evaluation-representative real machine.

Closing the metric bottleneck (train-only re-scoring). Decomposing the AdaSub ensemble gives source-AUC 0.7396, target-AUC 0.6456, pAUC 0.5995: pAUC and the 10-clip target domain are the bottleneck. Three train-only re-scorings of the *same* autoencoders—no retraining—address them: per-domain mean centring of both the residual Mahalanobis and the reconstruction-energy term (replacing the global mean $\bar{e}$ by the per-domain $\bar{e}_d$, removing the systematic target-reconstruction bias), shrinkage of the noisy target covariance toward the source, and P_{95} rather than P_{90} frame aggregation (sharper in the pAUC tail). Each is robust under the paired ten-seed test and together they reach Ω_{dev}=0.6544 (source-AUC 0.7441, target-AUC 0.6607, pAUC 0.6037; target $+0.0151$, pAUC $+0.0042$, source $+0.0045$ over the AdaSub ensemble).

What does not help (robustly). Several components strong at one seed do not survive the ten-seed protocol. An M2D

Table 3: Per-machine and overall development Ω_{dev} (harmonic mean of the three official terms) for the four submitted ten-seed ensembles. The champion (S1, AdaSub+DM+shrink+P_{95}) reaches 0.6544, best overall; the dev-Ω-selected cross-front-end ensemble S4 matches it (0.6534), while the coherence-leaning S3 trades valveEmu (≈ 0.52) for the real fan (0.660). Baseline per-machine values are harmonic means from [6], whose 21-term mean is the official 0.5766.

System	ToyCar	ToyCarEmu	bearingEmu	fan	gearboxEmu	sliderEmu	valveEmu	Overall
Baseline MAHALA [6]	.613	.624	.628	.518	.589	.543	.543	0.5766
S1 champion (AdaSub+DM+shr+P_{95})	.700	.645	.614	.626	.655	.585	.796	**.6544**
S2 RefSub (gentle denoise)	.715	.652	.615	.617	.641	.585	.607	.6308
S3 Combo (hedged fan-bet)	.713	.606	.613	**.660**	.650	.572	.523	.6140
S4 Dev-top10 (cross front-end)	.713	.647	.616	.610	.647	.581	**.808**	.6534

Table 4: Per-machine median G and per-machine Ω_{dev} for AdaSub vs CohGate; $\star$ is the only CohGate win. Ensembles: AdaSub 0.647, label-oracle 0.657.

Machine	G	AdaSub	CohGate
ToyCar	44.1	.698	.680
ToyCarEmu	2.20	.639	.575
bearingEmu	1.28	.599	.599
fan $\star$	0.94	.602	**.654**
gearboxEmu	0.92	.645	.634
sliderEmu	1.14	.582	.577
valveEmu	1.21	.774	.516

outlier-exposure scorer [10, 9] gives $+0.011$ at the anchor seed but -0.008 ± 0.008 over ten; LoRA [21] and adapters on frozen encoders stay within noise (best $+0.001$); a discriminative gate strong only under *test-set* normalisation (0.640–0.644) is guard-rejected (train-only LOMO $0.629<0.632$); training-objective changes are harmful ($-0.11/-0.08\,\Omega_{\text{dev}}$); and from-scratch normalizing-flow and bottleneck-kNN fusions show no robust signal (target-AUC collapses on the 10-clip target). Every robust gain is thus a train-only re-scoring, motivating a deliberately simple, pretrained-free system.

Per-machine transfer and the coherence diagnostic. Across the candidate front-ends (Table 2), CohGate is the most aggressive front-end, raising the fan-only score the most (0.6551) but lowering all-dev. Table 4 reports, per machine, the noise transfer G and the scores of AdaSub and CohGate. As the table shows, the only machine CohGate wins is fan; a label-oracle that picks the better front-end per machine reaches an ensemble of only 0.6571, just $+0.0101$ over AdaSub's 0.6470. Moreover G cannot isolate fan (G_{fan}=0.94 exceeds $G_{gearboxEmu}$=0.92, yet AdaSub wins gearbox), so the rule is not separable and the coherence gain is specific to one real machine. Crucially, the five evaluation machines all lie in the low-G fan regime ($G \in [0.74, 2.08]$, near the low-G end of the development range; Table 4), the most favourable data for the front-end.

Submitted systems. Four ten-seed ensembles are submitted and compared per machine in Table 3. The champion (S1) improves over the baseline on six of seven machines—largest on valveEmu ($+0.25$), fan ($+0.11$) and ToyCar ($+0.09$)—with a slight bearingEmu regression. The four span a denoising-aggressiveness axis plus a selected ensemble: (S1) the champion (AdaSub + per-domain-mean centring of the Mahalanobis and energy terms, target-covariance shrinkage, P_{95}; Ω_{dev}=0.6544); (S2) RefSub, a gentler denoise under the same stack (Ω_{dev} 0.6308); (S3) Combo, a guard-compliant per-clip blend of AdaSub and CohGate (fan-regime bet; 0.6140); and (S4) a dev-Ω-selected top-10 ensemble of the highest-scoring per-seed models across the AdaSub/RefSub front-ends (0.6534, chosen on development only, the evaluation set never seen). Every term is per-clip and train-only, so the pipeline transfers unchanged: front-ends, G, domain means, shrinkage and gamma are re-estimated on the evaluation normals while global scalars stay frozen.

5. CONCLUSION

This report presented a noise-aware front-end for first-shot ASD that uses the far microphone as a train-only noise reference (minimum-statistics noise transfer, floored adaptive spectral subtraction, Mahalanobis-scored autoencoder, ten-seed ensemble), plus three train-only post-hoc re-scorings targeting the pAUC/target bottleneck (per-domain-mean centring of the Mahalanobis and energy terms, target-covariance shrinkage, P_{95}). Under a strict multi-seed guarded protocol it lifts the official development score from 0.5766 to 0.6470 and then 0.6544 without retraining, alongside clean negative results (outlier exposure, architecture ablations, training-objective changes, and from-scratch fusion). As the evaluation set is entirely real, two-channel, and low-G, the gain is expected to transfer; future work will pursue learned reference denoisers.

6. ACKNOWLEDGEMENT

This research was supported by "Establishment of Evaluation Platform for Noise Reduction Technologies in Advanced Automotive Vehicles" through the Korea Institute for Advancement of Technology (KIAT) funded by the Ministry of Trade, Industry and Resources (MOTIR) (No. P0031293) and by the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

7. REFERENCES

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[2] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection under domain shifted conditions," in *Proc. DCASE Workshop*, 2021.

[3] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE Workshop*, 2022.

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

[5] 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 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE2024 Challenge, Tech. Rep., 2024.

[6] 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 preprint arXiv:2606.01578*, 2026.

[7] N. Harada, D. Niizumi, *et al.*, "DCASE 2023 task 2 baseline auto-encoder system (dcase2023t2-ae)," https://github.com/nttcslab/dcase2023_task2_baseline_ae, 2023.

[8] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 816–820.

[9] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. ICLR*, 2019.

[10] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Towards a universal audio pre-training framework," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2391–2406, 2024.

[11] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2023.

[12] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[14] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[15] 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.

[16] 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. DCASE Workshop*, 2022, pp. 1–5.

[17] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, and S. Vernekar, "A robust deep autoencoder for anomaly detection via mahalanobis distance," *arXiv preprint arXiv:1812.02765*, 2018.

[18] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1979, pp. 208–211.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[20] N. Yousefian and P. C. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 599–609, 2012.

[21] 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 *Proc. ICLR*, 2022.