

When Speech is Noise: Speech Removal for Machine Anomaly Detection

Edo, Takahiro; Wichern, Gordon; Fujimura, Takuya; Boeddeker, Christoph; Masuyama, Yoshiki;
Richter, Julius; Saijo, Kohei; Le Roux, Jonathan

TR2026-162 October 29, 2026

Abstract

Machine anomaly detection is widely used for predictive maintenance. When anomalous sound detection (ASD) systems are deployed in environments such as factories, we have noticed that a large number of false alarms are caused by human speech and nonverbal sounds such as laughing. In this paper, we investigate using a controlled evaluation framework where we contaminate existing ASD datasets with human vocalizations under various signal-to-noise ratios and durations. Experiments on different ASD systems show that human vocalizations significantly degrade anomaly detection performance. We further demonstrate that training with vocalization-contaminated data and applying source separation as a front-end both mitigate this degradation and partially restore anomaly-score distributions towards uncontaminated machine behavior.

Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) 2026

© 2026 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

When Speech is Noise: Speech Removal for Machine Anomaly Detection

Takahiro Edo^{1,2}, Gordon Wichern¹, Takuya Fujimura^{1,3}, Christoph Boeddeker¹,
Yoshiki Masuyama¹, Julius Richter¹, Kohei Saijo^{1,2}, Jonathan Le Roux¹

¹Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

²Mitsubishi Electric Corporation, Japan ³Nagoya University, Nagoya, Japan

Abstract—Machine anomaly detection is widely used for predictive maintenance. When anomalous sound detection (ASD) systems are deployed in environments such as factories, we have noticed that a large number of false alarms are caused by human speech and nonverbal sounds such as laughing. In this paper, we investigate using a controlled evaluation framework where we contaminate existing ASD datasets with human vocalizations under various signal-to-noise ratios and durations. Experiments on different ASD systems show that human vocalizations significantly degrade anomaly detection performance. We further demonstrate that training with vocalization-contaminated data and applying source separation as a front-end both mitigate this degradation and partially restore anomaly-score distributions towards uncontaminated machine behavior.

Index Terms—anomalous sound detection, speech removal, source separation, machine sound analysis

1. INTRODUCTION

Anomalous sound detection (ASD) for machine devices has become a core technology for predictive maintenance and condition monitoring in industrial facilities [1]. In the typical unsupervised setup, models are trained on only normal operating data, while at inference time, the deviation from the normal model is used as an anomaly score. An ASD task [2] was introduced as a Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Challenge, and subsequent editions have addressed increasingly realistic conditions [3]–[7]. This has enabled benchmarking of important algorithms such as autoencoders [8]–[11], generative models [12]–[14], and embedding-based approaches [15]–[18] among others, which are quickly becoming suitable for real-world deployment.

When ASD systems are deployed in practical environments such as factories, machines often operate in the presence of human workers. Unlike the background noise handled in the ASD task in the DCASE challenges, which is typically contained throughout the recordings, speech and nonverbal sounds such as laughing or shouting sporadically contaminate machine sounds. We collectively refer to these as *human vocalizations* in this paper. These human vocalizations can introduce a domain mismatch between models trained on machine sounds and contaminated mixtures encountered during deployment, leading to degraded detection performance and increased false alarms. Therefore, ASD systems need to handle these sporadic human vocalizations while preserving performance on clean machine sounds.

This motivates the fundamental question in this paper of whether general source separation models [19]–[21], which have become impressively powerful in the deep learning era, can be used as front-ends to suppress human vocalizations for ASD systems.

However, source separation or speech enhancement is not always beneficial as a front-end for downstream tasks. For example, artifacts from speech enhancement models are known to often degrade automatic speech recognition performance [22]–[24]. Separation artifacts have also been studied in the context of human listening [25], [26] and soundtrack analysis [27]. However, it remains unclear whether separation artifacts may affect ASD performance when machine sounds are mixed with human vocalizations.

In this paper, we construct, using publicly available data, a controlled evaluation that simulates contamination across signal-to-noise ratio (SNR) ranges and vocalization duration conditions. We investigate two strategies for improving ASD robustness: training the ASD model with human vocalization contamination and applying source separation as a front-end. Our experiments show that both strategies partially restore anomaly-score distributions toward uncontaminated machine behavior. However, because source separation operates entirely at inference time, it offers a practical means of improving robustness without requiring changes to the ASD training pipeline. We will release scripts to reproduce the dataset and benchmark.¹

2. BACKGROUND

2.1. Anomalous sound detection

Given a single-channel machine sound signal $m \in \mathbb{R}^T$ of length T , an ASD system produces an anomaly score $\mathcal{A}_\theta(m) \in \mathbb{R}$ with model parameters θ learned only on normal operating data. An anomaly is declared if $\mathcal{A}_\theta(m)$ is greater than a threshold τ . In this paper, we consider two representative ASD models implemented in the open-source toolkit ASDKit² [28].

Autoencoder (AE), the long-running official DCASE Challenge baseline [2]–[7], is trained to minimize the mean squared error (MSE) loss between the input and reconstructed log-mel spectrograms on normal data. Specifically, we use the `ae` recipe in the toolkit, which trains an independent AE for each machine type. During inference, the reconstruction error is used as the anomaly score $\mathcal{A}_\theta$.

BEATs-based discriminative method uses BEATs [29], a pre-trained audio foundation model, as the backbone. We adopt the discriminative method implemented in ASDKit [28]. The model is trained as a discriminative front-end by classifying meta-information labels associated with normal training data, such as machine type and machine ID. Specifically, we use the `dis_beats_scac_trainable` recipe in the toolkit, which fine-tunes BEATs with LoRA. In this recipe, the sequence of features output by BEATs is further aggregated and projected into a fixed-dimensional embedding. At test time, anomaly scores are computed in a distance-based manner: the embedding of a test sample is compared with embeddings of normal training samples. In our setting, we adopt a k-nearest neighbor (kNN)-based approach [16] and use the cosine distance to the nearest normal embedding as the anomaly score $\mathcal{A}_\theta$.

2.2. Task-aware Unified Source Separation (TUSS)

TUSS [20] takes as input a single-channel mixture $x \in \mathbb{R}^T$, and N prompts that indicate the source classes to extract. Unlike separation models where the number of output sources is fixed, the number of TUSS outputs depends on N . The prompts come from a set of fixed class labels, which can be combined to solve

¹<https://github.com/merlresearch/asd-speech-removal>

²<https://github.com/TakuyaFujimura/dcase-asd-toolkit>

many existing or new separation tasks, e.g., [speech, sfxbg] for speech enhancement, [speech, speech] for two-speaker speech separation, or [drums, bass, vocals, other] for typical four-stem music separation [30].

We use TUSS in this study for several reasons. First, the model was trained entirely on publicly available data, and the pretrained models and source code are open source³ (we employ the pretrained “medium” model). Second, TUSS is trained using an SNR-based loss, which explicitly encourages phase and scale alignment between the separated output and the input mixture, preserving potentially important signal characteristics for downstream anomaly detection. Finally, training a source separation model specifically for machine sounds would require clean machine recordings, which are not readily available. We believe the multi-task nature of TUSS can help to overcome this, as training on music has been shown to improve spontaneous speech separation [31] and may also help separate machine sounds due to shared transient and harmonic structure.

We also considered the recent generative separation model SAM-audio [21], but found the regenerated audio was too different from the mixture for downstream anomaly detection.

3. TASK AND DATA SETUP

3.1. Proposed pipeline

When deploying ASD systems, we have observed false alarms caused by human vocalizations. To improve robustness, we investigate two approaches: training the ASD model with speech-contaminated machine sounds and using a source separation front end that estimates vocalization-free machine sounds before applying a vanilla ASD model. Unlike mixture training, the separation approach requires no changes to the ASD or source separation models.

Let x denote an observed mixture consisting of a machine sound m and potentially a human vocalization h :

$$x = m + h. \quad (1)$$

A source separation model can be used to estimate the machine sound component $\hat{m}$, either directly or by first estimating the human vocalization component $\hat{h}_x$ from the mixture x and then subtracting it from the mixture:

$$\hat{m} = x - \hat{h}_x. \quad (2)$$

The estimated machine signal $\hat{m}$ is subsequently fed into an ASD model trained solely on machine sounds that have not been contaminated by human vocalizations.

3.2. Dataset building blocks

Machine sounds: We use the Development Dataset of the DCASE 2020 Challenge Task 2 [2], as this allows us to investigate performance without further complicating factors such as domain shifts introduced in more recent challenges. The dataset includes six machine types (fan, pump, slider, ToyCar, ToyConveyor, and valve), with recordings of 10 or 11 seconds in duration at 16 kHz sampling rate. It provides 2804 to 4000 normal sounds per machine for training. The test set contains both normal and anomalous sounds, with counts ranging from 400 to 2399 for normal samples and from 456 to 1475 for anomalous samples across machine types.

Human vocalizations: Speech is obtained from LibriSpeech [32], which contains clean read English speech from a single speaker per file. As speech samples vary widely in duration, we restrict our analysis to segments shorter than the 10-second machine recordings and divide

them into five duration bins: [0–2], [2–4], [4–6], [6–8], and [8–10] s. For each machine sound, five mixtures are generated by separately mixing it with one speech sample randomly selected from each of the five duration bins. For each mixture, the speech onset is uniformly sampled from the valid range so that the speech sample fits within the machine recording, without controlling speech activity.

Nonverbal sounds are extracted from FSD50K [33] using tags corresponding to nonverbal vocal events (e.g., laughter, screaming, shout, whispering, sigh, crying, and sobbing), following the categorization strategy adopted in DnR-nonverbal [34]. The selected clips span a wide range of durations. As with speech, one nonverbal sample from each duration bin is mixed with each machine recording, resulting in five mixtures for each machine recording.

3.3. Mixing protocol

In environments such as factories, the relative loudness of human vocalizations and machine sounds can vary substantially, so we generate mixtures across multiple target SNR ranges.

For each machine sound m and human vocalization h , we generate a mixture x by scaling h with a gain factor α to match a predefined target SNR. In this paper, the machine sound is treated as the target signal, while the human vocalization is regarded as additive noise. For each mixture, a target SNR value is randomly sampled from a predefined SNR range. The mixture is defined as

$$x = m + \alpha h, \quad (3)$$

where α is a scaling factor. Following previous sound mixing work [34]–[37], we define SNR using the difference in integrated loudness measured in Loudness Units relative to Full Scale (LUFS) [38]:

$$\text{SNR}_{\text{LUFS}} = \text{LUFS}(m) - \text{LUFS}(h). \quad (4)$$

To achieve a target SNR, the gain in decibels is

$$G_{\text{dB}} = \text{SNR}_{\text{LUFS}} - \text{SNR}_{\text{target}}, \quad (5)$$

and converted to a linear scaling factor:

$$\alpha = 10^{G_{\text{dB}}/20}. \quad (6)$$

We use three ranges for $\text{SNR}_{\text{target}}$: $[-9, 0]$, $[-3, 6]$, and $[3, 12]$ dB.

ASD mixture training: We replace the training set with speech-contaminated mixtures generated by mixing each normal training machine sound with a speech sample of duration [4–6] s and an SNR sampled from $[-3, 6]$ dB. Nonverbal sounds and other SNR and duration values are used to evaluate generalization performance.

4. EXPERIMENTS

4.1. Evaluation Protocol

ASD Metrics: As a representative example, Fig. 1 shows the anomaly score distributions computed using the BEATs-based model, with overlaid smooth probability density curves estimated using Kernel Density Estimation. Here, $\hat{h}_m$ and $\hat{h}_x$ denote the human vocalization estimates from m and x , respectively; Est(original) and Est(speech) are $m - \hat{h}_m$ and $x - \hat{h}_x$.

As shown in the contaminated mixture case (2nd row), both NORMAL and ANOMALY samples exhibit increased anomaly scores, resulting in a noticeable rightward shift of the overall distribution compared to the original machine sound condition. After applying source separation to the mixture (bottom row), the anomaly score distributions shift back toward the distribution observed for the original machine sound. This suggests that source separation effectively mitigates the effect of contamination and reduces the score shift caused

³<https://github.com/merlresearch/unified-source-separation>

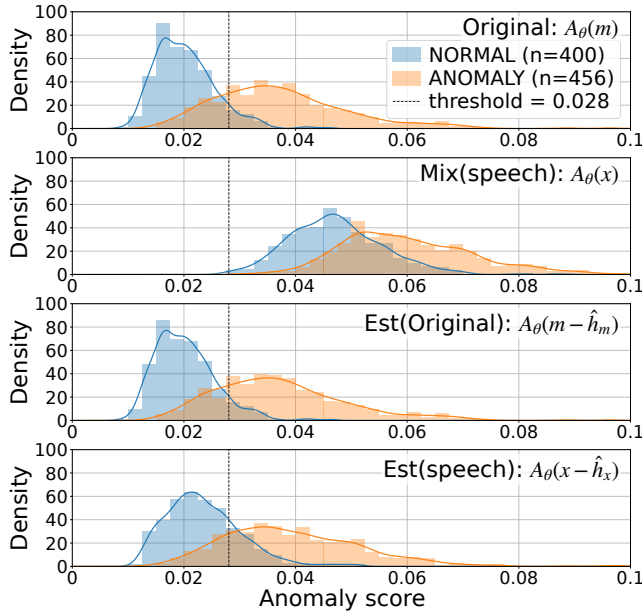


Fig. 1: Comparison of anomaly score distributions for the pump machine with speech contamination and front-end speech removal.

by domain mismatch. In the following, we introduce quantitative metrics to evaluate these effects.

Typically, the area under the curve (AUC) and partial AUC (pAUC) are used to evaluate ASD performance. These metrics are used as the official ranking metrics in the DCASE 2020 Challenge [2], [28]. Here, AUC denotes the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR), while pAUC denotes the partial area under the ROC curve computed over the low-FPR range $0 \leq \text{FPR} \leq 0.1$. Although AUC and pAUC are standard metrics in ASD, they can be difficult to interpret when human vocalization contamination causes a domain shift that increases the anomaly scores of both normal and anomalous machine sounds, as shown in Fig. 1 (2nd row). Because AUC and pAUC mainly reflect the relative ranking between normal and anomalous samples, they may not fully capture the score shift caused by contamination. Therefore, in addition to AUC and pAUC, we analyze TPR and FPR at a fixed threshold to evaluate how human vocalization contamination affects anomaly scores and detection decisions.

Since practical ASD systems require a decision threshold to determine whether a recording is anomalous, and selecting an appropriate threshold is challenging, in this paper, we determine an “oracle” threshold using the uncontaminated test data based on the Youden index [39]. This threshold is used only for analysis to evaluate how contamination affects the anomaly scores relative to the uncontaminated condition. The Youden index is defined as

$$J(\tau) = \text{TPR}(\tau) - \text{FPR}(\tau), \quad (7)$$

and the optimal threshold $\tau_{\max}$ is selected as:

$$\tau_{\max} = \arg \max_{\tau} J(\tau). \quad (8)$$

In Fig. 1, the dashed vertical line indicates the Youden-based threshold calculated using the uncontaminated machine sounds. This threshold is then used to compute TPR and FPR for contaminated and speech-removed data in the following quantitative evaluation. We report the average across machine types to evaluate ASD performance.

Table 1: Comparison of separation performance using SI-SNRi and output SI-SNR, with an average input SI-SNR of 8.2 dB before separation over the evaluated mixtures.

Prompt	Eval.	SI-SNRi		Output SI-SNR
		Mix (speech)	Mix (nonverb)	Original
SE	Eq. (2)	9.8	5.8	46.4
SE	sfxbg	9.1	2.8	37.7
C	Eq. (2)	9.7	5.1	47.3
MS	Eq. (2)	9.3	8.9	21.4

Table 2: Comparison of anomaly detection performance averaged over machine types, with duration [4–6] s and SNR [−3,6] dB.

Input	Autoencoder (%)				BEATs (%)			
	AUC↑	pAUC↑	TPR↑	FPR↓	AUC↑	pAUC↑	TPR↑	FPR↓
Original	72.3	59.5	71.2	35.1	89.5	79.9	80.9	14.6
Est (original)	72.3	59.4	70.6	35.1	89.5	79.8	80.8	15.2
Mix (speech)	60.6	53.4	98.0	92.9	79.1	66.1	97.7	78.7
Est (speech)	65.5	54.8	91.4	73.2	81.8	68.4	93.1	55.3
Mix (nonverb)	61.5	53.6	97.0	90.0	77.5	62.7	95.9	69.0
Est (nonverb)	64.3	53.7	93.0	76.6	81.0	65.1	92.9	53.4

Separation Metrics: As shown in Fig. 1, when source separation is applied to the original machine sound without contamination (3rd row), the resulting distribution remains largely unchanged compared to the original condition. This suggests that artifacts introduced by the separation model have a limited impact on the anomaly score distribution when the input consists of clean machine sounds.

To quantitatively evaluate both the effectiveness of removing human vocalizations and the artifacts introduced by the source separation model, we use scale-invariant signal-to-noise ratio (SI-SNR) [40], with the original machine sound as the reference. Input SI-SNR is computed from the mixture, output SI-SNR from the separated machine sound, and SI-SNR improvement (SI-SNRi) is defined as their difference.

When the original machine sound is used as the input to the source separation system, the input SI-SNR is infinite because no contamination is present. Therefore, the evaluation is performed using only the output SI-SNR instead of improvement.

TUSS Configurations: To evaluate the performance of removing human vocalizations from machine sounds, we provide TUSS with three different prompt sets: speech enhancement (SE), music separation (MS), and a combination (C) prompt set. The SE and MS prompt sets are derived from the prompts described in Section 2.2. The C prompt set consists of classes taken from both prompt sets (e.g., [speech, sfxbg, drums, other]), a combination that we found to work well in initial experiments. In all configurations, the machine sound estimated as residual by Eq. (2) is used for evaluation. For the SE prompt set, which contains only [speech, sfxbg], we also consider using the sfxbg output directly as the estimated machine sound, leading to four configurations in total. Finally, we report the average SI-SNRi or SI-SNR across all conditions for each prompt combination.

4.2. Results

Table 1 shows separation performance for the four prompt configurations described above. When comparing results for the SE prompt set, we observe that treating the estimated sfxbg signal as the machine sound degrades performance. This result indicates that subtracting the estimated speech from the mixture, as in Eq. (2), is most effective. This is non-trivial as TUSS does not constrain the output sources to sum up to the mixture, and seems to indicate that TUSS can more robustly estimate speech than the more diverse sfxbg class.

Table 3: Comparison of anomaly detection performance averaged over machine types under baseline, source separation, and mixture training.

Test condition			Autoencoder (%)									BEATs (%)								
			baseline			+ source separation			+ train mixture			baseline			+ source separation			+ train mixture		
Type	Dur.	SNR	AUC↑	pAUC↑	FPR↓	AUC↑	pAUC↑	FPR↓	AUC↑	pAUC↑	FPR↓	AUC↑	pAUC↑	FPR↓	AUC↑	pAUC↑	FPR↓	AUC↑	pAUC↑	FPR↓
Original	–	–	72.3	59.5	35.1	72.3	59.4	35.1	68.1	56.8	39.5	89.5	79.9	14.6	89.5	79.8	15.2	86.7	76.6	13.2
Speech	[4–6]	[−3,6]	60.6	53.4	92.9	65.5	54.8	73.2	62.3	53.5	87.6	79.1	66.1	78.7	81.8	68.4	55.3	86.8	75.9	11.0
Nonverb	[4–6]	[−3,6]	61.5	53.6	90.0	64.3	53.7	76.6	61.2	53.2	86.5	77.5	62.7	69.0	81.0	65.1	53.4	86.5	75.3	12.3
Speech	[0–2]	[3,12]	69.2	56.6	60.0	70.6	57.1	48.3	67.4	56.0	53.1	84.3	70.6	38.5	85.7	72.4	31.3	87.0	76.7	12.0
Speech	[0–2]	[−9,0]	64.1	55.3	89.3	66.8	55.3	71.7	65.3	55.0	81.1	81.7	67.1	57.7	85.3	73.6	41.7	87.4	76.9	11.5
Speech	[8–10]	[−9,0]	54.9	51.1	100.0	59.5	52.1	94.7	56.9	52.0	99.8	66.3	55.7	100.0	74.6	62.5	89.7	83.0	70.6	30.6
Speech	[8–10]	[3,12]	61.5	54.4	86.4	66.3	55.1	67.5	62.3	54.0	83.2	76.1	63.6	82.8	80.1	65.6	56.5	85.8	74.7	16.8

For mixtures containing speech contamination, the SE prompt set performs best, with the C prompt set (which is a superset of SE) obtaining comparable performance. Interestingly, for mixtures containing nonverbal sounds, the MS prompt set performs best, possibly because the `vocals` prompt is better matched to nonverbal vocalizations. However, this prompt set tends to degrade the signal quality of the original machine sounds, which may negatively affect anomaly detection performance.

Based on these observations, we use the C prompt set for the subsequent anomaly detection experiments because it provides high SI-SNR while causing relatively small degradations to the original machine sounds.

Table 2 summarizes the anomaly detection performance averaged over machine types with duration [4–6] s and SNR [−3,6] dB. Compared with Original, Mix degrades AUC and pAUC, while Est(speech) and Est(nonverb) partially recover these values. When comparing TPR and FPR, Mix shows much higher values for both metrics than Original, as the speech contamination causes nearly all samples to be labeled anomalous. In contrast, Est(speech) and Est(nonverb) reduce both TPR and FPR compared with Mix and move them toward Original. This tendency is consistently observed for both ASD models.

Table 3 compares three settings: the baseline setting without preprocessing or contaminated training data (“baseline”), source separation preprocessing by TUSS before ASD (“+ source separation”), and training with speech-contaminated mixtures (“+train mixture”). The first three columns specify the test condition: vocalization type, duration, and SNR. The test set includes original machine sounds, matched speech and nonverbal mixtures with duration [4–6] s and SNR [−3,6] dB, and four mismatched speech conditions using duration [0–2] or [8–10] s and SNR [−9,0] or [3,12] dB. Overall, the FPR tends to increase as the SNR range decreases and the vocalization duration increases. This tendency was consistently observed across different ASD models, indicating that longer and stronger human vocalization contamination tends to increase anomaly scores and thus leads to more false positives. Source separation consistently reduced the FPR compared with the baseline setting for both ASD models, but the FPR remained higher than that of the original machine sounds, suggesting that further improvements are needed to fully mitigate the impact of human vocalization contamination. For the BEATs-based model, training with contaminated mixtures substantially reduces the FPR across all conditions, suggesting that mixture training is more effective than source separation preprocessing when a BEATs-based ASD model can be retrained.

5. DISCUSSION

The results show that contamination by human vocalizations increases anomaly scores and consequently leads to higher TPR and FPR even for normal machine sounds. Source separation partially mitigates this score shift and moves the score distribution toward the uncontaminated

condition. This suggests that source separation can serve as an effective preprocessing step when an existing ASD model must be used without retraining.

For the BEATs-based model, training with speech-contaminated mixtures substantially reduces FPR, suggesting that BEATs can improve robustness when the ASD model can be retrained. In contrast, for the autoencoder model, mixture training is less effective than source separation, indicating that simply replacing clean normal data with contaminated mixtures does not necessarily improve reconstruction-based ASD.

These observations suggest that source separation and mixture training are suited to different deployment scenarios. Source separation is model-agnostic and can be applied to existing ASD systems, whereas mixture training can be more effective when contaminated mixture data are available and the ASD model can be retrained. However, this strategy requires meta-information labels for discriminative training and can increase the size of the neighbor memory used for distance-based scoring, or require an additional strategy for selecting representative neighbors, which can increase system complexity.

On the other hand, several limitations remain in our separation-based approach. First, the nonverbal category used in this study does not cover all possible types of nonverbal sounds. Moreover, although the TUSS prompt sets selected in this paper do not introduce significant artifacts into uncontaminated machine sounds, other source separation models may introduce artifacts that could affect ASD performance. In addition, speech and nonverbal contamination were evaluated separately in our experiments. In real-world environments, however, these contamination sources may co-occur and overlap in more complex ways, which could further increase the difficulty of anomaly detection. In this initial study, we did not jointly optimize the separation and ASD models. Furthermore, this study primarily focuses on contamination caused by human vocalizations. In practical environments, machine sounds may also be affected by other human noises such as footsteps or background activities, which may be more difficult to separate from machine sounds. Recent DCASE ASD challenges have also considered general background noise.

6. CONCLUSION

In this paper, we investigated ASD under realistic contamination by human vocalizations. Using a controlled mixture protocol, we quantified the robustness degradation caused by both speech and nonverbal interference. Experimental results showed that removing human vocalization components using a source separation model mitigates anomaly-score shifts and partially restores score distributions toward those of clean machine sounds. We also found that training with contaminated mixtures can improve robustness when retraining is possible. These findings highlight the importance of contamination-aware evaluation and suggest that front-end source separation and mixture training are promising approaches for improving the robustness of ASD systems in real industrial environments.

REFERENCES

- [1] Y. Kawaguchi and T. Endo, “How can we detect anomalies from subsampled audio signals?” in *Proc. IEEE MLSP*, Sep. 2017, pp. 1–6.
- [2] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda *et al.*, “Description and discussion on DCASE2020 Challenge Task2: Unsupervised anomalous sound detection for machine condition monitoring,” in *Proc. DCASE*, 2020, pp. 81–85.
- [3] 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 detection for machine condition monitoring under domain shifted conditions,” in *Proc. DCASE*, 2021, pp. 186–190.
- [4] 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*, 2022.
- [5] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, “Description and discussion on DCASE 2023 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring,” in *Proc. DCASE*, 2023.
- [6] 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,” in *Proc. DCASE*, 2024, pp. 111–115.
- [7] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit *et al.*, “Description and discussion on DCASE 2025 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring,” in *Proc. DCASE*, 2025, pp. 55–59.
- [8] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, “Group masked autoencoder based density estimator for audio anomaly detection,” in *Proc. DCASE*, 2020, pp. 51–55.
- [9] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, “Anomalous sound detection based on interpolation deep neural network,” in *Proc. IEEE ICASSP*, 2020, pp. 271–275.
- [10] G. Wichern, A. Chakrabarty, Z. Wang, and J. Le Roux, “Anomalous sound detection using attentive neural processes,” in *Proc. IEEE WASPAA*, 2021, pp. 186–190.
- [11] K. Shimonishi, K. Dohi, and Y. Kawaguchi, “Anomalous sound detection based on sound separation,” in *Proc. ISCA Interspeech*, 2023, pp. 2733–2737.
- [12] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, “Unsupervised detection of anomalous sound based on deep learning and the neyman-pearson lemma,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 212–224, 2019.
- [13] T. Fujimura, K. Imoto, and T. Toda, “Discriminative neighborhood smoothing for generative anomalous sound detection,” in *Proc. EUSIPCO*, 2024, pp. 156–160.
- [14] J. Yin, Y. Gao, W. Zhang, T. Wang, and M. Zhang, “Diffusion augmentation sub-center modeling for unsupervised anomalous sound detection with partially attribute-unavailable conditions,” in *Proc. IEEE ICASSP*, 2025.
- [15] H. Chen, Y. Song, Z. Zhuo, Y. Zhou, Y.-H. Li, H. Xue, and I. McLoughlin, “An effective anomalous sound detection method based on representation learning with simulated anomalies,” in *Proc. IEEE ICASSP*, 2023.
- [16] P. Saengthong and T. Shinozaki, “Deep generic representations for domain-generalized anomalous sound detection,” in *Proc. IEEE ICASSP*, 2025.
- [17] A. Jiang, X. Zheng, B. Han, Y. Qiu, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, “Adaptive prototype learning for anomalous sound detection with partially known attributes,” in *Proc. IEEE ICASSP*, 2025.
- [18] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, “Local density-based anomaly score normalization for domain generalization,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 33, pp. 4642–4652, 2025.
- [19] J. Pons, X. Liu, S. Pascual, and J. Serra, “GASS: Generalizing audio source separation with large-scale data,” in *Proc. IEEE ICASSP*, 2024, pp. 546–550.
- [20] K. Saijo, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, “Task-aware unified source separation,” in *Proc. IEEE ICASSP*, Apr. 2025.
- [21] B. Shi, A. Tjandra, J. Hoffman, H. Wang, Y.-C. Wu, L. Gao, J. Richter, M. Le, A. Vyas, S. Chen, C. Feichtenhofer, P. Dollár, W.-N. Hsu, and A. Lee, “SAM Audio: Segment anything in audio,” *arXiv preprint arXiv:2512.18099*, 2025.
- [22] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, T. Moriya, and N. Kamo, “Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition,” in *Proc. ISCA Interspeech*, 2021, pp. 1149–1153.
- [23] Y. Koizumi, S. Karita, A. Narayanan, S. Panchapagesan, and M. Bacchiani, “Snri target training for joint speech enhancement and recognition,” in *Proc. ISCA Interspeech*, 2022, pp. 1173–1177.
- [24] T. Ochiai, K. Iwamoto, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, “Rethinking processing distortions: Disentangling the impact of speech enhancement errors on speech recognition performance,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 3589–3602, 2024.
- [25] M. Torcoli, J. Paulus, T. Kastner, and C. Uhle, “Controlling the remixing of separated dialogue with a non-intrusive quality estimate,” in *Proc. IEEE WASPAA*, 2021, pp. 91–95.
- [26] N. L. Westhausen, R. Huber, H. Baumgartner, R. Sinha, J. Rennie, and B. T. Meyer, “Reduction of subjective listening effort for tv broadcast signals with recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3541–3550, 2021.
- [27] D. Petermann, G. Wichern, A. S. Subramanian, Z.-Q. Wang, and J. Le Roux, “Tackling the cocktail fork problem for separation and transcription of real-world soundtracks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2592–2605, 2023.
- [28] F. Takuya, W. Kevin, I. Keisuke, and T. Tomoki, “ASDKit: A toolkit for comprehensive evaluation of anomalous sound detection methods,” in *Proc. DCASE*, 2025.
- [29] 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.
- [30] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “MUSDB18-HQ - an uncompressed version of MUSDB18,” Dec. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.3338373>
- [31] S. Uhlich, G. Fabbro, M. Hirano, S. Takahashi, G. Wichern, J. Le Roux, D. Chakraborty, S. Mohanty, K. Li, Y. Luo, J. Yu, R. Gu, R. Solovyev, A. Stempkovskiy, T. Habruseva, M. Sukhovei, and Y. Mitsufuji, “The Sound Demixing Challenge 2023 – Cinematic Demixing Track,” *Trans. ISMIR*, Apr. 2024.
- [32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.
- [33] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, “FSD50K: An open dataset of human-labeled sound events,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2022.
- [34] T. Hasumi and Y. Fujita, “DnR-Nonverbal: Cinematic audio source separation dataset containing non-verbal sounds,” in *Proc. ISCA Interspeech*, 2025.
- [35] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, “Scaper: A library for soundscape synthesis and augmentation,” in *Proc. IEEE WASPAA*, 2017, pp. 344–348.
- [36] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, “WHAM!: Extending speech separation to noisy environments,” in *Proc. ISCA Interspeech*, 2019.
- [37] D. Petermann, G. Wichern, Z.-Q. Wang, and J. Le Roux, “The cocktail fork problem: Three-stem audio separation for real-world soundtracks,” in *Proc. IEEE ICASSP*, 2022.
- [38] E. Grimm, R. Van Everdingen, and M. J. L. C. Schöpping, “Toward a recommendation for a european standard of peak and LKFS loudness levels,” *SMPTE Motion Imaging Journal*, vol. 119, no. 3, pp. 28–34, 2010.
- [39] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
- [40] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?” in *Proc. IEEE ICASSP*, 2019, pp. 626–630.