2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Spectral-to-Spatial Distillation: Denoising Framework for Real-Time Anomalous Sound Detection

Koki Shoda†, Jun Younes Louhi Kasahara†, Takuya Igaue†,
Shinji Kanda†, Hajime Asama†, Qi An†, and Atsushi Yamashita†

***Abstract*— In this paper, we propose spectral-to-spatial distillation, a novel denoising framework for real-time anomalous sound detection. While anomalous sound detection is crucial for industrial applications, its reliability is often compromised by background noise, which can lead to false positives. Our proposed method addresses the issue of background noise by distilling knowledge from a general-purpose *spectral* filtering network into an environment-specific *spatial* filtering network. Specifically, we generate distillation targets, which are audio signals with reduced noise, using a pre-trained foundation model. A spatial filtering network is then trained using these targets. A key feature of our distillation process is its ability to automatically generate these targets using only one-shot, brief, noise-free reference signal of the target sound. Furthermore, we introduce a new quality metric for these distillation targets, called Semantic Clarity improvement (SCi). By leveraging the semantic audio embedding capabilities of a foundation model, SCi measures the improvement in semantic similarity between the distillation target and the reference signal. This SCi allows for effective distillation by weighing the loss function based on the quality of the targets. Experimental results demonstrate that our method achieves the best denoising and anomaly detection performance while maintaining real-time processing capabilities, making it a practical solution for noisy industrial environments.**

I. Introduction

In large-scale industrial facilities such as manufacturing factories and oil refineries, where a vast number of machines operate simultaneously, the use of artificial intelligence for automatic anomaly detection is of paramount importance for maintenance and operational safety. While visual anomaly detection using images plays a significant role in this domain [1], anomaly detection based on acoustic signals is equally crucial, as it can non-intrusively identify internal machine failures that are not visible [2]. By placing microphones at key locations within these facilities, it is possible to monitor the sounds of surrounding machinery and detect anomalies.

Anomalous sound detection is typically approached as a one-class classification problem, where models such as One-Class SVM [3] or Autoencoders [4] learn the distribution of normal sound data. Anomalies are then identified as deviations from this learned distribution. This approach is adopted because, in real-world factory settings, actual anomalous sounds occur rarely and exhibit great diversity, making it impractical to collect or intentionally create a comprehensive dataset of anomalous sounds. However, a challenge arises from the difficulty of distinguishing between machine anomalous sounds and ambient background noise, which often leads to false positives. Therefore, effective background noise reduction is a prerequisite for robust acoustic anomaly detection.

Corresponding author: Koki Shoda (shoda@robot.t.u-tokyo.ac.jp).
† The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.

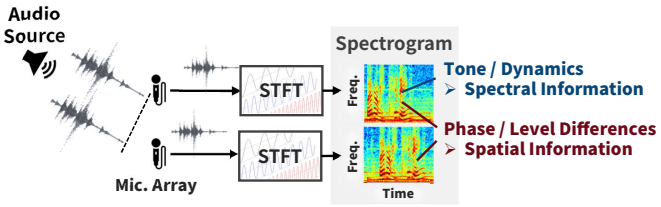


Fig. 1: Comparison between *spectral* and *spatial* information. STFT stands for the short-time fourier transform. A spectrogram is a two-dimensional representation of audio signals, where the horizontal axis represents time, the vertical axis represents frequency, and the color represents the amplitude of the signal.

In scenarios where multi-channel audio signals from microphone arrays are available, two types of information can be leveraged for denoising: *spectral* information and *spatial* information [5], as illustrated in Fig. 1. Spectral information, obtainable from a single channel, contains content-related details such as tone and dynamics. In contrast, spatial information is formulated from inter-channel phase and level differences [6], capturing sound propagation and sound field characteristics through constructs such as steering vectors and inter-channel phase differences.

A. Spectral Information for Denoising

Numerous supervised learning-based methods have been proposed for denoising using only spectral information [7]. These spectral filtering methods, often employing time-series models such as Transformers [8] or RNNs [9], are trained to extract a target sound from a noisy audio mixture. However, training is typically limited to normal sounds. A model trained to extract normal sounds is ill-suited for denoising anomalous sounds, as it may treat the anomaly itself as noise to be removed.

An alternative approach, query-based source separation, offers a solution by allowing the target to be specified dynamically. The most common form is language-queried separation, where models such as AudioSep [10] can extract sounds based on text descriptions in a zero-shot manner. The main strength of this method, its reliance on natural language, also becomes its critical weakness in industrial environments. Text queries such as “machinery” often lack the specificity to distinguish one machine from many others, leading to poor separation performance.

978-1-6654-5784-2/26/$31.00 ©2026 IEEE

To resolve this ambiguity, Audio-Queried Source Separation (AQSS) provides a more precise solution by using a reference audio clip as the query. For industrial applications, this is ideal, as a clean recording of a machine's normal sound can precisely define the target. Despite this advantage, current AQSS methods, such as VoiceFilter [11], introduce a new bottleneck: they require extensive fine-tuning on large, paired datasets for every new target sound. This need for a costly, per-target training process renders them impractical for large-scale industrial deployment. Thus, the key challenge is to create an AQSS system that retains the precision of an audio query without the heavy requirement of fine-tuning.

B. Spatial Information for Denoising

Blind Source Separation (BSS) [12], [13] have been proposed for denoising using spatial information. BSS leverages the statistical independence and spatial properties of sound sources to separate them without prior knowledge of their direction or timbre. However, BSS typically requires an iterative optimization process, often involving dozens of iterations through auxiliary function-based methods [14], [15], to solve for the source models and separation filters for each inference. Consequently, its computation time often exceeds the duration of the audio being processed. This makes true real-time inference, defined as the ability to process a segment of audio in a time shorter than its duration (e.g., processing a one-second chunk in less than one second), unfeasible. This is a critical limitation for machine anomaly detection, as any system that cannot process audio faster than it is recorded will inevitably fail in a continuous monitoring scenario due to accumulating data.

Additionally, spatial filtering neural networks [16], [17], [18] have been also proposed for denoising using spatial information. These methods do not use spectral features directly; instead, they learn to extract the target sound source based solely on frequency-dependent spatial cues, such as inter-channel level and phase differences. Assuming that both normal and anomalous sounds originate from the same machine, their location remains constant. Consequently, their spatial signature will be identical, regardless of the sound's content. Therefore, a spatial filter trained on the spatial signature of normal sounds should be equally effective at extracting anomalous sounds from the same source, as the model is agnostic to the spectral content of the signal. However, training these networks requires a large amount of paired data (clean target sound and mixed sound), which cannot be captured simultaneously in a live environment. Consequently, a practical training strategy for real-world deployment has not been established.

C. Objective

As discussed, existing methods present critical trade-offs. Spectral approaches risk suppressing the very anomalies they aim to detect, require impractical per-target fine-tuning, or rely on ambiguous language queries. Spatial approaches, while powerful, are often too slow for real-time use or lack a practical training strategy without clean ground-truth data. To address these multifaceted challenges, the objective of this research is to develop a denoising framework for anomaly detection that is both real-time capable and operates based on audio cues, thereby eliminating the dependency on language queries. To achieve this objective, we propose a framework that synergistically utilizes both spectral and spatial information for denoising.

Our contributions are summarized as follows:

- **Fine-Tuning-Free Audio-Queried Source Separation:** We propose a method for Audio-Queried Source Separation (AQSS) that requires no fine-tuning by combining existing foundation models. Existing AQSS models [11] often require training to extract sounds similar to a reference audio clip from a mixture. Thus, we propose a novel approach of generating distillation targets, which are denoised audio signals used as training targets for a supervised spatial filtering network. Crucially, this process requires only a single reference audio clip of the target sound to generate the distillation targets.
- **Ground-Truth-Free Quality Assessment for Effective Distillation:** Training with distillation targets, which inevitably contain more residual noise compared to ideal clean audio, poses a significant challenge. A student model trained naively on these targets may simply learn to replicate their imperfections. To address this, we introduce Semantic Clarity improvement (SCi), a novel metric for assessing the quality of generated signals without requiring ground-truth clean audio. SCi leverages a foundation model to quantify the semantic similarity improvement between the separated audio and the reference audio. The SCi allows for effective loss weighting, enabling the student model to learn robustly even from noisy distillation targets.
- **High-Fidelity Denoising with Real-Time Capability:** Our proposed framework achieves both high-fidelity denoising performance and real-time inference capability. By distilling knowledge from a pretrained spectral filter into an environment-specific spatial filter, we overcome the common trade-off between performance and speed, delivering a truly practical solution for industrial deployment.

II. Related Work

Knowledge distillation is a technique where a compact *student* model is trained to replicate the output of a larger, more complex *teacher* model, often to achieve faster inference [19]. In the context of multi-channel audio processing, this paradigm can be applied to bridge different model types.

A conceivable application is 'spatial-to-spectral' distillation [20], [21]. Here, a computationally expensive but effective multi-channel method such as BSS [12], [13], which relies on spatial information, could act as the teacher. Its separated output could be used to train a lightweight, single-pass *student* network that operates only on spectral features. The primary motivation for such a transfer would be to distill the separation capability of a slow spatial algorithm into a much faster spectral filtering network for real-time applications.

To the best of our knowledge, our research proposes the reverse and novel direction of 'spectral-to-spatial' distillation. We use a large, general-purpose spectral filtering network as the *teacher* to train an environment-specific spatial filtering network as the *student*. We transfer the high-level denoising capability of a spectral filter to a spatial filter, which is

inherently better suited to preserve the characteristics of anomalous sounds originating from a fixed location. This approach uniquely addresses the limitations of both domains: it overcomes the tendency of spectral filters to suppress anomalies and bypasses the need for large-paired training datasets typically required for spatial filters, thus creating a practical and high-performance solution.

III. Proposed Method

A. Concept

Spectral and spatial information present a trade-off between generalization and the need for specialized training. Spectral information is readily available from single-microphone recordings, and large-scale, general-purpose denoising models [10] exist. However, due to the high variability of spectral patterns, their denoising performance can be limited in specific, noisy environments. Conversely, spatial information provides powerful cues for highly effective denoising, but it is heavily dependent on the specific microphone count, array geometry, and acoustic environment. This specificity means that generalized spatial models do not exist, and a new model must be trained for each deployment scenario.

To resolve this trade-off, we propose a novel framework that uses a spectral-based model to generate distillation targets for training a spatial filtering neural network, as shown in Fig. 2. In Fig. 2, HTS-Audio Transformer is the Hierarchical Token-Semantic Audio Transformer [22] and utilizes frozen weights from Contrastive Language-Audio Pretraining (CLAP) [23] while Residual U-Net is the source separation model based on spectral information [24] and utilizes frozen weights from AudioSep [10]. Narrow-Band LSTM is the spatial filtering neural network [16] and its weights are optimized by the distillation targets after random initialization. By training a model that relies solely on spatial cues to replicate the output of a spectrally-informed model, we effectively distill the denoising capability from the general-purpose spectral domain to an environment-specific spatial filter. This creates a fast and high-performance denoiser without requiring ground-truth clean data from the target environment.

A spatial filtering network is an excellent choice for anomaly detection denoising for the following reasons:

- **Denoising Performance on Anomalous Sounds:** The network learns to extract a source based only on its spatial signature (inter-channel level and phase differences). Assuming that both normal and anomalous sounds originate from the same machine and the surrounding acoustic environment is stable, their spatial characteristics will be identical. Therefore, a filter trained on normal sounds will be equally effective at extracting anomalous sounds.
- **Fast Inference:** Denoising is achieved in a single forward pass through the network, enabling real-time inference. This is a stark contrast to BSS [12], [13] methods, which require computationally expensive iterative optimization, making them orders of magnitude slower.

For our spatial filtering network, we employ a Narrow-Band Long Short-Term Memory (LSTM) [16]. While other architectures such as Transformers [17], [18] have been proposed for spatial filtering, LSTMs possess a strong inductive bias for sequential data. This allows them to achieve robust performance with fewer parameters, which not only facilitates faster inference but also mitigates the risk of overfitting to the potentially imperfect distillation targets generated by the spectral filter.

B. Mathematical Framework

Our proposed method consists of two parallel processing streams: a distillation target generation stream using spectral information and a student model training stream using spatial information. Given a multi-channel noisy mixture signal $\boldsymbol{m}(t) = [m_1(t), m_2(t), \ldots, m_C(t)]$ captured by a C-channel microphone array, we first transform it into the time-frequency domain via the STFT:

$$\boldsymbol{M} = \text{STFT}[\boldsymbol{m}(t)] \in \mathbb{C}^{C \times T \times F}, \tag{1}$$

where T and F denote the number of time frames and frequency bins, respectively.

The one-shot reference audio $y_{\text{one-shot}}(t)$ is encoded into a semantic embedding vector $\boldsymbol{c}$ using the HTS-Audio Transformer $f_{\text{HTS-AT}}$:

$$\boldsymbol{c} = f_{\text{HTS-AT}}(\text{Mel-FB}[\text{STFT}[y_{\text{one-shot}}(t)]]; \boldsymbol{\theta}_{\text{HTS-AT}}), \tag{2}$$

where $\boldsymbol{\theta}_{\text{HTS-AT}}$ are frozen parameters pretrained by CLAP [23], and Mel-FB denotes the Mel-filter bank operation. Conditioned on the encoded reference $\boldsymbol{c}$, the spectral filtering network $f_{\text{U-Net}}$ extracts the target sound from the first microphone channel $\boldsymbol{M}(0)$, producing a separation magnitude mask $|\boldsymbol{E}|$ and a phase residual $\angle\boldsymbol{R}$:

$$|\boldsymbol{E}|, \angle\boldsymbol{R} = f_{\text{U-Net}}(\boldsymbol{M}(0), \boldsymbol{c}; \boldsymbol{\theta}_{\text{U-Net}}), \tag{3}$$

where $\boldsymbol{\theta}_{\text{U-Net}}$ are frozen parameters pretrained by AudioSep [10]. The distillation target spectrogram $\boldsymbol{Y}_{\text{distill}}$ is then constructed as follows:

$$\boldsymbol{Y}_{\text{distill}} = |\boldsymbol{E}| \odot \boldsymbol{M}(0) \odot e^{j\angle\boldsymbol{R}}, \tag{4}$$

where $\odot$ denotes element-wise multiplication.

In parallel, the spatial filtering network $f_{\text{NB-LSTM}}$ processes the full multi-channel mixture $\boldsymbol{M}$ to estimate the target sound spectrogram $\hat{\boldsymbol{Y}}$:

$$\hat{\boldsymbol{Y}} = f_{\text{NB-LSTM}}(\boldsymbol{M}; \boldsymbol{\theta}_{\text{NB-LSTM}}). \tag{5}$$

The trainable parameters $\boldsymbol{\theta}_{\text{NB-LSTM}}$ are optimized to minimize the discrepancy between its output and the distillation target in the time domain:

$$\mathcal{L}\{\text{iSTFT}[\hat{\boldsymbol{Y}}], \text{iSTFT}[\boldsymbol{Y}_{\text{distill}}]\}, \tag{6}$$

where iSTFT denotes the inverse STFT operation, and $\mathcal{L}$ is a weighted loss function detailed in Eq. (7).

C. Audio-Queried Source Separation without Fine-Tuning

To generate distillation targets, we introduce a method for AQSS that requires no fine-tuning. Our core strategy is to adapt an existing language-queried model, AudioSep [10], to accept audio queries instead. We achieve this by replacing AudioSep's text encoder with the corresponding audio encoder from its underlying CLAP [23]. This simple yet effective modification allows us to use a reference audio clip

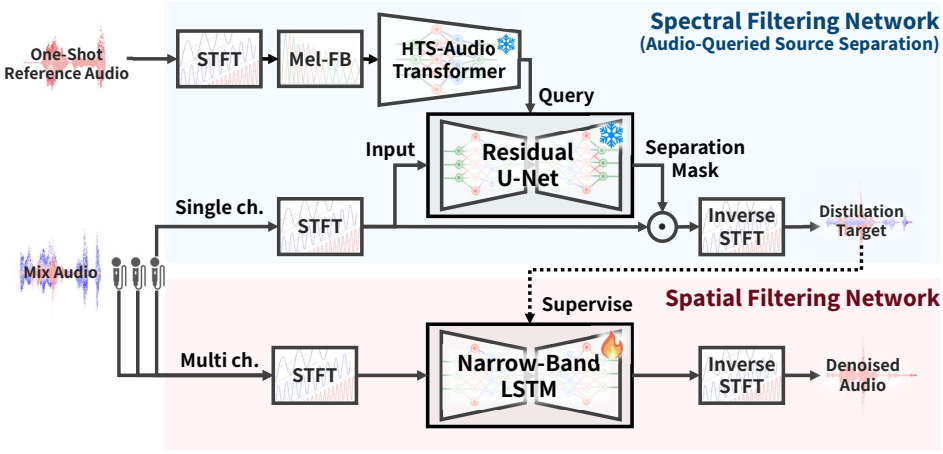


Fig. 2: Strategy of distilling spectral-based model into spatial filtering neural network and overall architecture of the proposed method. The flame icon indicates modules that are optimized by the distillation targets, while the snowflake icon indicates modules whose parameters are frozen. Here, $\odot$ denotes the element-wise multiplication. By training a model that relies solely on spatial cues to replicate the output of a spectrally-informed model, we effectively distill the denoising capability from the general-purpose spectral domain to a specialized, high-performance spatial filter.

as a query instead of text, thereby realizing a fine-tuning-free AQSS system.

This AQSS approach, combining the power of CLAP and AudioSep, can generate a denoised sound source (a distillation target) from just a single clean audio sample of the target machine. This dramatically reduces the data requirement, minimizing the need for extensive, hard-to-acquire clean recordings. Furthermore, this approach bypasses the ambiguity and manual effort associated with crafting precise text queries, making the system more robust and easier to deploy.

This direct replacement is made possible by the specific training strategy of AudioSep. The model builds upon CLAP, which consists of two encoders, an audio encoder and a text encoder, trained jointly to align their respective embedding spaces. Crucially, the parameters of the text encoder are frozen during the training of AudioSep. This preserves the learned correspondence between the text and audio embedding spaces, which is the key property that enables our audio-queried approach to work without any retraining.

D. Loss Weighting by Semantic Clarity Improvement

Distillation targets generated using only spectral information have limited denoising quality. Spatial filtering networks are conventionally trained on pairs of clean audio and noisy mixtures. Since our distillation targets inevitably contain more residual noise than true clean audio, a naive training approach may limit the network's performance to that of the distillation target.

Therefore, we propose to weight the training loss based on the quality of the distillation targets. However, conventional quality metrics such as Signal-to-Noise Ratio improvement (SNRi) require access to the ground-truth clean audio, which is unavailable in our scenario. To circumvent this, we propose a new metric of Semantic Clarity improvement (SCi) as a proxy for quality.

As illustrated in Fig. 3, SCi leverages the audio encoder of CLAP [23] to measure the improvement in semantic similarity between the distillation targets and the one-shot reference audio. We then use this SCi score to weight the primary training objective: the Scale-Invariant Signal-to-Noise Ratio (SI-SNR) between the network's output and the distillation target. This allows us to effectively down-weight or filter out low-quality distillation targets, facilitating more robust and effective training. The training loss function for the spatial filtering network is defined as:

$$\mathcal{L}\{\hat{y}(t), y_{\text{distill}}(t)\} = \mathbb{E}\{\mathcal{S}[\text{SCi}] \cdot \text{SI-SNR}[\hat{y}(t), y_{\text{distill}}(t)]\}, \quad (7)$$

where $\mathbb{E}$ denotes the expectation, $\hat{y}(t)$ is the network's output at time t, $y_{\text{distill}}(t)$ is the distillation target at time t, SI-SNR is the Scale-Invariant Signal-to-Noise Ratio, and $\mathcal{S}$ is a robust scaling function that clips SCi values by percentile and scales them to the range $[0, 1]$. The SI-SNR is defined by reference audio $a(t)$ and estimated audio $\hat{a}(t)$ as follows:

$$\text{SI-SNR}[\hat{a}(t), a(t)] = 10 \log_{10} \frac{\|\delta a(t)\|^2}{\|\delta a(t) - \hat{a}(t)\|^2}, \quad (8)$$

$$\delta = \frac{\langle \hat{a}(t), a(t) \rangle}{\|a(t)\|^2}, \quad (9)$$

where the notation $\langle \hat{a}(t), a(t) \rangle$ denotes the inner product of the $\hat{a}(t)$ and $a(t)$.

IV. Experiments

To evaluate the effectiveness of the proposed method, we conducted experiments on a multi-channel dataset. To the best of our knowledge, no public dataset exists for anomalous sound detection in multi-channel environments with diverse, interfering noise sources. Therefore, we constructed a synthetic dataset by combining several existing single-channel audio datasets.

For the target machine sounds, we used the ToyADMOS dataset [25], which contains recordings of both normal and anomalous operating sounds. ToyADMOS is a benchmark for anomalous sound detection, featuring three classes of

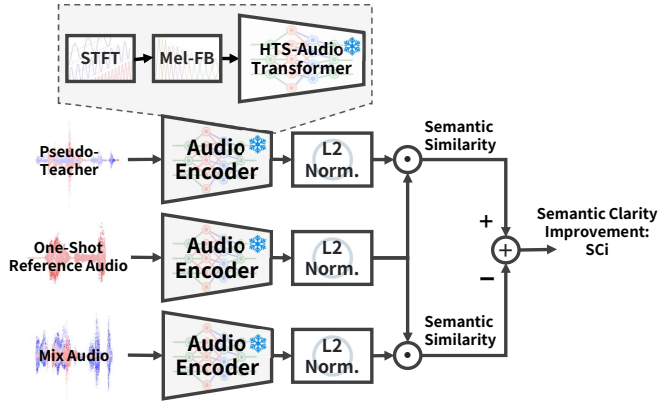


Fig. 3: Diagram of the SCi calculation. Here, $\odot$ denotes the inner product, which is used to calculate the semantic similarity to the one-shot reference audio. SCi quantifies how much the distillation target, generated via audio-queried source separation, becomes semantically closer to the one-shot reference audio as a result of denoising.

miniature machines: Car, Conveyor, and Train. Each audio clip is 10 seconds long and the dataset includes a variety of realistic anomalous types, such as shaft deformation, gear damage, under/over voltage, and foreign objects on a conveyor belt.

For the noise sources, we prepared four classes of sounds representing diverse acoustic scenes. Given that industrial facilities can be both indoors and outdoors, these noises include sounds commonly found in and around such environments: Mechanisms (e.g., machinery, vehicles), Speech (multilingual conversations), Music (diverse genres), and Animals (e.g., birds, insects from outdoor settings). The Mechanisms and Animals audio clips were sourced from the BSD10k dataset [26], while the Speech and Music clips were taken from the MUSAN dataset [27]. For all noise classes, we used all available audio clips that were 10 seconds or longer.

To create a realistic multi-channel environment, we synthesized spatial audio by convolving the source signals with real-world measured Room Impulse Responses (RIRs) from a calibrated 5-channel dataset [28]. These RIRs were recorded in a cuboid room measuring 6 m $\times$ 6 m $\times$ 2.4 m with a reverberation time of 150 ms, capturing acoustic characteristics from seven sound source locations and six microphone array positions [28]. For our experiments, the target sound source location and the microphone array position were kept fixed. The noise source, however, was randomly assigned to one of the three other available locations for each synthesized audio mixture. All audio was resampled to a 16 kHz sampling rate and trimmed to a duration of 10 seconds.

A. Training Dataset

To adhere to the practical constraints of anomaly detection, our training set consists exclusively of normal sounds as the target. To match the size of our evaluation set, we randomly selected 528 normal samples for each of the three machine classes from ToyADMOS. For our proposed method, which requires only a single clean reference, we designated just one of these samples per class as the clean reference audio. The remaining samples were mixed with noise at a SNR randomly chosen from a range of -10 to 0 dB. The noise was created by randomly selecting one of the four noise classes and then clipping a 10-second segment from a random file within that class. For training the baseline spectral filtering method [7], [8], [9], which requires supervised learning with clean-noisy pairs, we used a separate set of 256 clean target samples and their corresponding noisy mixtures.

B. Evaluation Dataset

The evaluation dataset was constructed to contain an equal number of normal and anomalous sounds. We used 264 anomalous sound samples for each of the three machine classes from ToyADMOS, this being the maximum number that allows for balanced classes. We then randomly selected an equal number of 264 normal sound samples, ensuring no overlap with the training set. All 528 samples (264 normal, 264 anomalous) in the evaluation set for each class were mixed with noise at a random SNR between -10 and 0 dB. Similar to the training dataset, the noise was generated by randomly selecting one of the four noise classes and then clipping a 10-second segment from it.

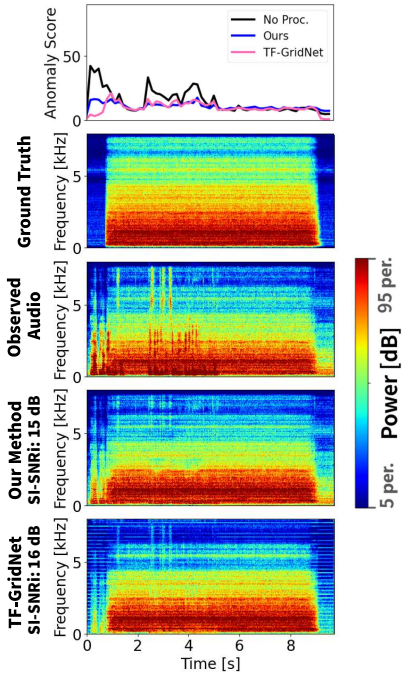
C. Implementation Details

For all trainable models (both our proposed method and the comparison methods), we used the Adam optimizer with a learning rate of 10^{-3}. Early stopping [29] was employed, halting the training process when the validation loss did not decrease for a set number of epochs. The validation set was created by randomly holding out 20% of the training data. All models were implemented in Python 3.10 using PyTorch, with the SSSPy library being used for BSS baselines. For the BSS baselines [12], [13], we set the number of bases to 8, a conventional value for this task. The iterative optimization process, which updates the source models and demixing filters, was run for 100 iterations to ensure convergence.

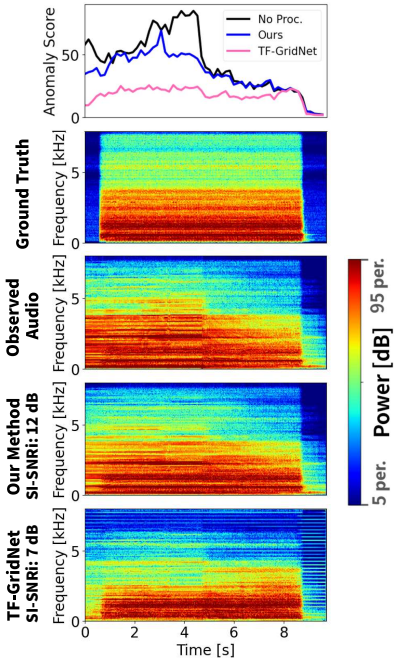
To ensure a fair comparison of computational speed, all inference times were measured on a single CPU (Intel Core i9-12900KS). We opted against GPU acceleration because all baseline methods, particularly the BSS algorithms, are not optimized for GPUs. To assess the performance of anomaly detection, we adopted the baseline feature extraction method and anomaly detector from the DCASE 2020–2025 Challenge [30]. For feature extraction, log-Mel spectrograms were computed using a 64 ms frame length and a 32 ms frame shift, resulting in 128 Mel bands. These features from five consecutive frames were then concatenated to form a 640-dimensional feature vector. The anomaly detector is an 8-layer Autoencoder, which was trained exclusively on the features of normal sounds to learn a compressed representation of normal operation. The final anomaly score is defined as the squared reconstruction error between the input features and the Autoencoder's output.

V. Results

Examples of the anomaly detection and denoising results are shown in Fig. 4. In Fig. 4 (a), the target sound is in a normal state, so a lower anomaly score is desirable. This example contains significant human speech in the first half. Without any processing, the anomaly score spikes in sync with the speech events. In contrast, both our proposed



(a) In this example, the *normal-state* Conveyor is the target sound, and noise from the Speech class is used. This example contains significant human speech in the first half. Without any processing, the anomaly score spikes in sync with the speech events.



(b) In this example, the *anomalous-state* Conveyor is the target sound, and noise from the Mechanisms class is used, which includes the operating sounds of a different machine in the first half. Our proposed method achieves an effective SI-SNRi of 12 dB, indicating substantial background noise reduction. Furthermore, it yields a higher anomaly score compared to the TF-GridNet [7].

Fig. 4: Example of the anomaly detection and denoising results. The top figures show the anomaly detection scores and the others show spectrograms. "Ground Truth" are the clean target sounds, and "Observed Audio" are the noisy mixture, which is the input to the denoising model.

method and the existing method (TF-GridNet [7]) successfully prevent this rise in the anomaly score, demonstrating that both are effective in preventing false positives caused by noise.

In Fig. 4 (b), the target sound is in an anomalous state, making a higher anomaly score desirable. In this example, noise from the Mechanisms class is used, which includes the operating sounds of a different machine in the first half. Our proposed method achieves an effective SI-SNRi of 12 dB, indicating substantial background noise reduction. Furthermore, it yields a higher anomaly score compared to the existing method (TF-GridNet [7]). This suggests that our method effectively suppresses noise while preserving the characteristic features of the anomaly, making it highly suitable for denoising in anomaly detection tasks.

Figure 5 shows the scatter plot of the SCi and the SI-SNRi of the training dataset. The Spearman's correlation coefficient between SCi and SI-SNRi is 0.42, indicating a moderate correlation, and all of the data with SCi greater than 0.2 have SI-SNRi greater than 0. This suggests that SCi can be used as a proxy for SI-SNRi, enabling robust training without requiring ground-truth clean audio.

The overall denoising and anomaly detection performance on the entire evaluation dataset is presented in Table I. As shown in the table, our proposed method achieves the highest performance in both denoising and anomaly detection compared to the existing methods. Compared to "Our

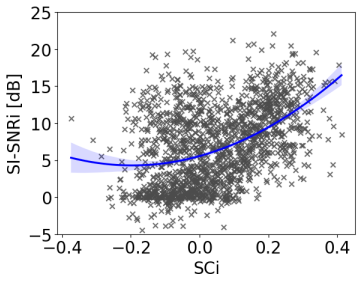


Fig. 5: Scatter plot of the SCi and the SI-SNRi. The blue line indicates the regression result and the blue band indicates the 95% confidence interval.

AQSS only", which generates the distillation targets, our full method demonstrates superior denoising and anomaly detection performance. This indicates that the denoising capability of the spectral-based AQSS is effectively distilled into the spatial filtering network. Notably, while spectral filtering-based methods exhibit high denoising performance (SI-SNRi), their anomaly detection accuracy can sometimes be even lower than that of "No processing". This seemingly contradictory result is further investigated in the Discussion section.

The average inference times are detailed in Table II. Among the BSS methods, ILRMA [13], being a separation

TABLE I: Performance of the individual methods. Denoising performance is evaluated by SI-SNRi (average ± the standard deviation). Anomaly detection performance is evaluated by Receiver Operating Characteristic Area Under the Curve (ROC-AUC). The best performance in each column is shown in **bold**.

Method	Approach Category	SI-SNRi [dB]			ROC-AUC		
		Car	Conveyor	Train	Car	Conveyor	Train
No processing	-	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.65	0.79	0.59
TF-GridNet [7]	Spectral Filtering (SF)	**8.8** ± 3.6	7.7 ± 5.4	13.0 ± 3.9	0.70	0.88	0.57
Conv-Transformer [8]	Spectral Filtering (SF)	6.0 ± 3.4	5.1 ± 4.4	10.4 ± 4.2	0.56	0.73	0.58
DualPathRNN [9]	Spectral Filtering (SF)	7.1 ± 3.4	5.9 ± 5.4	11.8 ± 3.9	0.73	0.86	0.62
AudioSep [10]	Language-Queried Source Separation (LQSS)	6.4 ± 6.5	4.0 ± 7.7	5.9 ± 10.1	0.71	0.81	0.55
Fast MNMF [12]	Blind Source Separation (BSS)	0.1 ± 0.9	0.4 ± 2.8	0.1 ± 1.3	0.63	0.77	0.59
ILRMA [13]	Blind Source Separation (BSS)	4.2 ± 3.3	3.6 ± 3.6	3.0 ± 3.0	0.73	0.86	0.62
Our AQSS only	Audio-Queried Source Separation (AQSS)	5.5 ± 5.1	6.1 ± 5.5	8.3 ± 5.9	0.73	0.89	0.59
Our Method	Spectral-to-Spatial Distillation (SSD)	8.3 ± 3.7	**8.4** ± 3.4	**13.6** ± 3.9	**0.75**	**0.91**	**0.64**

TABLE II: Inference time, expressed as the average ± the standard deviation, taken to process a 10-second audio sample. The best inference time is shown in **bold**.

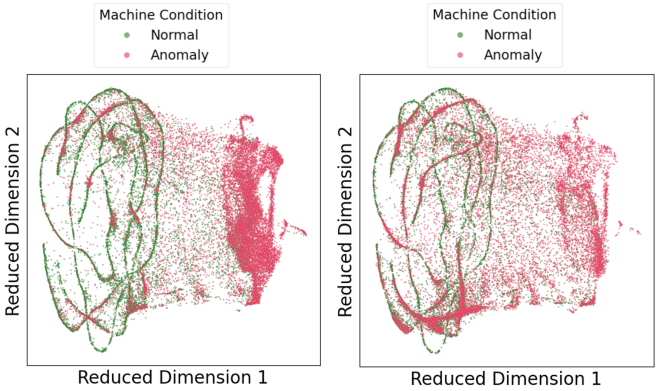
Method	App. Cat.	Inference Time [s]
TF-GridNet [7]	SF	10.56 ± 0.10
Conv-Transformer [8]	SF	0.55 ± 0.04
DualPathRNN [9]	SF	**0.24** ± **0.03**
AudioSep [10]	AQSS	2.42 ± 0.06
Fast MNMF [12]	BSS	45.69 ± 1.34
ILRMA [13]	BSS	6.51 ± 0.06
Our AQSS only	AQSS	2.00 ± 0.10
Our Method	SSD	0.53 ± 0.02

model-based approach, is faster than the generative model-based FastMNMF [12]. In our experimental setup, ILRMA's processing time was under 10 seconds, making it real-time capable for a 10-second audio clip. However, this is a borderline case. An increase in sampling rate or the use of a less powerful CPU would easily push the inference time beyond the 10-second threshold, rendering real-time processing infeasible. In sharp contrast, our proposed method processes the audio in approximately 1/20 of its real-time duration. This significant speed advantage ensures robust real-time performance across a wide range of hardware and scenarios, highlighting its practical applicability.

VI. Discussion

The core concept of our spectral-to-spatial distillation framework was validated as a successful approach. The strategy of distilling knowledge from a large, general-purpose spectral filter to train a specialized, efficient spatial filter proved highly effective for anomalous sound detection in noisy environments. Our experimental results comprehensively support this. As detailed in Table I, our framework achieved best performance in both denoising and anomaly detection, surpassing various established methods. This performance gain was achieved alongside exceptional computational speed suitable for real-time deployment, with the proposed SCi metric facilitating robust training from the generated distillation targets.

As observed in Table I, spectral filtering methods such as TF-GridNet, despite their high denoising performance, sometimes resulted in lower anomaly detection accuracy than with no processing at all. We hypothesized that this occurs because these models, trained only on normal sounds, learn to extract only the normal components from an anomalous sound, effectively treating the anomaly itself as noise to be removed.



(a) Feature distribution after denoising by our spatial filtering. Our spatial filtering-based method maintains a clear distinction between the normal and anomalous feature clusters.

(b) Feature distribution after denoising by the spectral filtering (TF-GridNet [7]). The spectral filtering method reveals that the features corresponding to anomalous sounds are incorrectly mapped into the cluster of normal sounds.

Fig. 6: Feature distributions after denoising in Train. The spectral filtering method tends to strip out the anomalous characteristics from the sound.

To investigate this hypothesis, we visualized the feature distributions after denoising using UMAP [31], a dimensionality reduction technique. The UMAP model was trained on the features of the original clean audio, labeled as normal or anomalous. The results are shown in Fig. 6.

As shown in Fig. 6 (a), our spatial filtering-based method maintains a clear distinction between the normal and anomalous feature clusters. In contrast, Fig. 6 (b) reveals that with the spectral filtering method, the features corresponding to anomalous sounds are incorrectly mapped into the cluster of normal sounds. This confirms that the supervised spectral filtering method tends to strip out the anomalous characteristics from the sound. Therefore, the high SI-SNRi achieved by this method can be attributed to the fact that the amount of suppressed noise was far greater than the amount of the desired signal that was discarded. While this improves the metric, it is detrimental to the task of anomaly detection.

A limitation of the current framework is its assumption of static sound sources, which may not always hold in real-world scenarios. Future work will address this by incorporating an online spatial filtering network [32] capable of tracking and adapting to moving sound sources in real-

time. Furthermore, to facilitate deployment on resource-constrained edge devices, we will explore using additional knowledge distillation techniques to compress the trained spatial filter into an even more computationally efficient model.

VII. Conclusion

We proposed spectral-to-spatial distillation, a novel denoising framework that bridges the gap between general-purpose spectral filtering networks and environment-specific spatial filtering networks. Our framework leverages a fine-tuning-free AQSS method to generate distillation targets from a single reference audio clip, effectively distilling knowledge from a large foundation model into a lightweight, environment-specific spatial filter. Experimental results demonstrated that our method achieves high-fidelity denoising and anomaly detection performance, outperforming existing spectral filtering and BSS approaches. Crucially, it accomplishes this with exceptional computational speed, enabling robust real-time inference on standard CPU hardware. Our framework thus achieves the objectives set in the introduction: a real-time, audio-cued denoising system that eliminates the dependency on language queries.

Acknowledgement

This work was supported in part by the Suzuki foundation. This work was also supported in part by the World-leading Innovative Graduate Study Program in Proactive Environmental Studies (WINGS-PES), The University of Tokyo.

References

[1] J. Liu, G. Xie, J. Wang, *et al.*, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024.

[2] 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 Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2019.

[3] Z. Mnasri, J. H. Giraldo, and T. Bouwmans, "Anomalous sound detection for road surveillance based on graph signal processing," in *Proceedings of the 2024 32nd European Signal Processing Conference*, 2024, pp. 161–165.

[4] J. Guan, Y. Liu, Q. Kong, *et al.*, "Transformer-based autoencoder with id constraint for unsupervised anomalous sound detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, 2023.

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation,," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[6] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization,," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, 2007.

[7] Z.-Q. Wang, S. Cornell, S. Choi, *et al.*, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[8] A. Gulati, J. Qin, C. C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of the 2020 Annual Conference of the International Speech Communication Association*, 2020, pp. 5036–5040.

[9] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 46–50.

[10] X. Liu, Q. Kong, Y. Zhao, *et al.*, "Separate anything you describe," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 458–471, 2025.

[11] Q. Wang, H. Muckenhirn, K. Wilson, *et al.*, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proceedings of the 2019 Annual Conference of the International Speech Communication Association*, 2019, pp. 2728–2732.

[12] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 12, pp. 2610–2625, 2020.

[13] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[14] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[15] A. Zaeemzadeh, M. Joneidi, N. Rahnavard, and M. Shah, "Iterative projection and matching: Finding structure-preserving representatives and its application to computer vision," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5409–5418.

[16] C. Quan and X. Li, "Multi-channel narrow-band deep speech separation with full-band permutation invariant training," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 541–545.

[17] C. Quan and X. Li, "Multichannel speech separation with narrow-band conformer," in *Proceedings of the 2022 Annual Conference of the International Speech Communication Association*, 2022, pp. 5378–5382.

[18] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.

[19] J. D. Geoffrey Hinton, Oriol Vinyals, "Distilling the knowledge in a neural network," in *Proceedings of the NIPS Workshop on Deep Learning and Representation Learning*, 2014, pp. 1–9.

[20] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, "Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 356–360.

[21] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 81–85.

[22] K. Chen, X. Du, B. Zhu, *et al.*, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 646–650.

[23] Y. Wu, K. Chen, T. Zhang, *et al.*, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[24] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2021, pp. 342–349.

[25] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 313–317.

[26] P. Anastasopoulou, J. Torrey, X. Serra, and F. Font, "Heterogeneous sound classification with the broad sound taxonomy and dataset," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2024, pp. 11–15.

[27] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," [Online] https://openslr.org/17/, 2015.

[28] D. Di Carlo, P. Tandeitnik, C. Foy, *et al.*, "Dechorate: a calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–15, 2021.

[29] L. Prechelt, "Automatic early stopping using cross validation: Quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998.

[30] Y. Koizumi, Y. Kawaguchi, K. Imoto, *et al.*, "Description and discussion on DCASE2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2020, pp. 81–85.

[31] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[32] C. Quan and X. Li, "Multichannel long-term streaming neural speech enhancement for static and moving speakers," *IEEE Signal Processing Letters*, vol. 31, pp. 2295–2299, 2024.