

基于滤波 – 锚稀疏自编码器的滑轨音频异常检测

彭 瑜^{1,2} 刘豪睿²
 PENG Yu LIU Haorui

摘 要

随着工业设备智能运维需求的提升，如何在缺乏故障标注的条件下实现高可靠的音频异常检测成为重要研究问题。文章以 MIMII 数据集中滑轨工况（6 dB 低噪场景）为主要研究对象，提出一种滤波 – 锚稀疏自编码器模型。实验在 MIMII 滑轨 6 dB 数据集上表明，相较多种典型自编码器，滤波 – 锚稀疏自编码器在 AUC 和 pAUC 指标上均获得提升，尤其在低假阳性率区间表现出更优的检测鲁棒性。研究结果表明，引入频域可学习滤波与稀疏锚注意力的自编码器框架，能够有效提升滑轨音频异常检测的精度与工程适用性。

关键词

异常检测；自编码器；音频信号；无监督学习

doi: 10.3969/j.issn.1672-9528.2026.01.017

0 引言

随着人工智能、智慧城市和物联网等技术的高速发展，社会逐渐步入数字化和自动化的阶段，自动化的异常声音检测也日益广泛地应用到工厂机器的日常安全监测工作中。但是异常声音种类繁多，存在难以收集且收集成本昂贵等难点。考虑到实际应用的特点，业界更多考虑的是通过无监督学习的方法来解决异常声音检测这项任务，即在训练集中不包含任何异常声音，只用正常声音来训练异常声音检测模型^[1]。

本文以 MIMII 数据集中滑轨工况为主要研究对象^[2]，提出一种滤波 - 锚稀疏自编码器（filter-anchor sparse

autoencoder, FA-SAE）用于无监督声学异常检测。该模型构建在注意力自编码器框架之上，首先在频域引入可学习滤波模块，对输入 Mel 频谱进行任务自适应的频带重加权，以突出与设备运行状态高度相关的关键频段、抑制冗余或噪声成分；随后在解码器施加锚点式稀疏约束，使少量“锚特征”集中编码正常工况下的主导声学模式，从而在重构误差空间中放大异常样本与正常样本的区分度。

总体而言，本文旨在结合频域滤波增强与锚式稀疏编码，构建面向滑轨工况的轻量化无监督声学异常检测模型，并在 MIMII 滑轨数据集上系统评估其在 AUC 与 pAUC 指标下的性能表现。通过将模型输出映射为对异常程度更为敏感的重构误差与评价指标，期望为工业滑轨等关键部件的早期故障预警提供具有工程可落地性与部署友好的声学监测方案。

1. 吉林化工大学信息与控制工程学院 吉林吉林 132022
 2. 德州学院计算机与信息学院 山东德州 253023

[2] 孙崇智, 吴永伟, 安建民, 等. 基于模糊 T-S 型内模 PID 控制算法的无刷直流电机仿真分析 [J]. 现代电子技术, 2024, 47(24): 18-24.
 [3] 古刚. 基于模糊 PID 的无刷直流电机控制方法 [J]. 电气技术与经济, 2025(2): 223-225.
 [4] 王光瑞, 李国洪, 张晓阳, 等. 基于模型预测的无刷直流电机控制研究 [J/OL]. 天津理工大学学报, 1-7[2025-12-18]. <https://link.cnki.net/urlid/12.1374.N.20251013.1838.016>.
 [5] NGUANG S K, ASSAWINCHAICHOTE W, SHI P. Robust H_{∞} control design for fuzzy singularly perturbed systems with Markovian jumps: an LMI approach[J]. IET control theory & applications, 2007, 4(1): 893-908.

【作者简介】

张灿（1998—），女，山东临沂人，硕士研究生，助教，研究方向：随机切换系统、模糊控制，email: zhangcan19981224@sina.com。
 朱晓飞（1987—），男，山东德州人，博士研究生，讲师，研究方向：智能制造、数字孪生、数字化设计与仿真，email: zxfei1987@126.com。
 胡红伟（1990—），男，河南周口人，硕士研究生，讲师，研究方向：深度学习、模糊控制，email: hongwei_hu20170706@alu.hit.edu.cn。

（收稿日期：2025-07-03 修回日期：2025-12-30）

1 相关技术

1.1 滑动窗口的 log-Mel 频谱

在工业音频异常检测中,前端特征提取在很大程度上决定了后续模型可分辨的模式空间,因此多数现有工作更倾向于先从原始波形构建稳健的时-频表示,而不是直接在时域信号上进行端到端建模。尽管亦有部分研究探索了基于原始波形的端到端建模方案,但在工业场景中,基于 log-Mel 频谱等时-频特征的做法仍然较为常见^[3]。针对本研究的机器声音数据,首先对每段 wav 信号进行短时傅里叶变换,并通过 Mel 滤波组获得维度为 (n_{mels}, T) 的 Mel 频谱图,再经对数幅度压缩形成 log-Mel 频谱,以减弱能量尺度差异和长尾分布的影响。在此基础上,沿时间轴施加长度为 frames、步长为 1 的滑动窗口,将相邻 frames 帧的 Mel 子块视作局部视频片段,并按列展平为维度为 $(n_{\text{mels}} \times \text{frames})$ 的定长向量^[4]。这样既保留了短时时间上下文与频率结构,又将变长音频统一映射到固定维空间,便于后续自编码器及一类检测模型进行训练与重构误差度量,如图 1 所示。

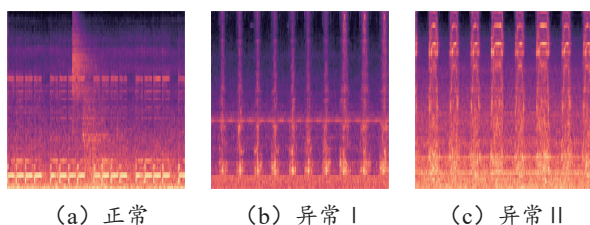


图 1 音频信号频谱提取后的特征

1.2 自编码器

无监督异常检测方法通常通过生成式或表示式模型学习正常数据的特征分布,将显著偏离该分布的样本判定为异常^[5]。自编码器 (Auto-Encoder) 是一种成熟的深度学习单分类网络方法,其以无监督的方式学习目标数据特征,只学习正常样本特征即可获得检测模型,适用于异常情况复杂多样性的异常检测。算法模型包含 1 个 Encoder (编码器) 和 1 个 Decoder (解码器): 编码器将高维输入编码成低维的隐变量从而强迫神经网络学习最优信息量的特征;解码器的作用是将隐藏层的隐变量还原到初始维度,最好的状态是解码器的输出能够完美地或者近似恢复出原来的输入^[6]。

在工业音频异常检测中,基于自编码器的无监督方法通常以若干帧对齐的时-频特征(如 log-Mel 频谱块)为输入,将局部时-频结构压缩到紧凑潜在表示,再通过解码器重构,以刻画设备在正常工况下的稳态声学模式及微小波动,其重构误差常被用作异常度量^[7]。

2 算法改进设计

2.1 PairFilter 的引入

为提升特征对设备状态相关谱结构的刻画能力,本文在

编码器输入端引入可学习滤波模块,如图 2 中 Filter 块^[8],以数据驱动的方式对滑动窗口 log-Mel 频谱块进行自适应谱整形。

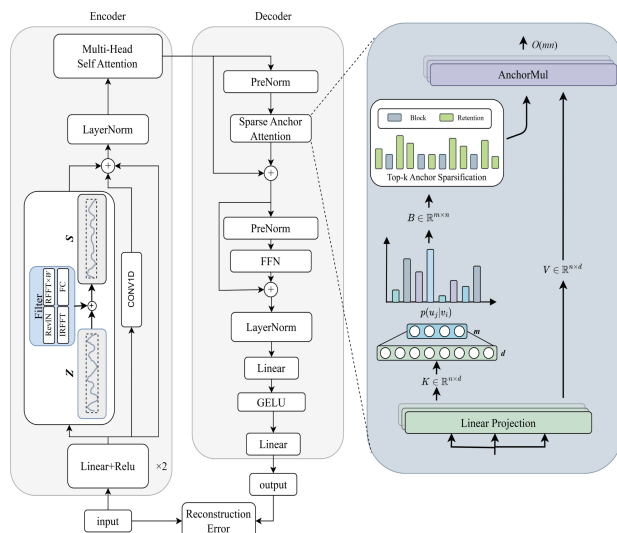


图 2 FA-SAE 模型构型

设输入特征为 $X \in \mathbb{R}^{T \times F}$, 其中 T 和 F 分别表示时间帧数与 Mel 频带数,学习型滤波模块基于当前样本生成上下文相关的滤波权重 $H_\theta(X) \in \mathbb{R}^{T \times F}$, 通过逐点加权实现频域选择性增强与抑制:

$$\tilde{X} = X \odot H_\theta(X) \quad (1)$$

式中: $\odot$ 表示逐元素乘法,经滤波后的特征 $\tilde{X}$ 作为自编码器的输入,有效抑制背景噪声与无关频带能量,突出与工况变化密切相关的谱形信息,从而在低噪声工况下获得更平滑、更可分的正常样本流形。

2.2 锚注意力机制及改进

锚注意力 (Anchor Attention)^[9] 是一类基于“锚点”的稀疏注意力形式,通过引入少量全局锚向量对序列进行摘要,从而在降低复杂度的同时保留关键全局结构。本文在自编码器的解码器结构中引入锚注意力。设编码器输出为 $Q \in \mathbb{R}^{N \times d}$, 其中 N 表示时间-频率位置的总数(例如将时频展平后的序列长度), d 表示每个位置的特征维度。通过选取或学习得到锚点集合 $A \in \mathbb{R}^{N \times M}$, M 为锚点数,首先计算样本到锚点的注意力权重:

$$B = \text{softmax}\left(\frac{QA^T}{\sqrt{d}}\right), B \in \mathbb{R}^{N \times M} \quad (2)$$

式中: $QA^T \in \mathbb{R}^{N \times M}$ 为每个位置与各锚点之间的相似度; $\sqrt{d}$ 为缩放因子; softmax 为每一行上归一化,得到位置对锚点的归一化关联强度。随后利用锚点对全局信息进行聚合与重投影:

$$Z = BA \in \mathbb{R}^{N \times d}, H = f_{\text{FFN}}(Z) \quad (3)$$

式中： $\mathbf{Z}$ 表示通过锚点汇聚后的全局上下文特征； $f_{\text{FFN}}(\cdot)$ 表示逐位置的前馈网络；输出 $\mathbf{H} \in \mathbb{R}^{N \times d}$ 表示锚注意力层的结果。

在锚注意力的基础上，本文进一步引入 Top-k 稀疏化策略^[10]，图2中 Sparse Anchor Attention 模块，以抑制注意力权重在锚点之间的过度扩散，并凸显与结构性模式最相关的少数关键锚点。设锚注意力权重矩阵为 $\mathbf{B} \in \mathbb{R}^{N \times M}$ ； $\mathbf{B}_{ij}$ 表示第*i*个位置对第*j*个锚点的注意力权重^[11]。

$$\hat{\mathbf{B}}_{ij} = \begin{cases} \mathbf{B}_{ij}, & j \in \text{Top-}k \\ 0, & \text{otherwise} \end{cases}, \mathbf{B}_{ij}^{(s)} = \frac{\hat{\mathbf{B}}_{ij}}{\sum_{l=1}^M \hat{\mathbf{B}}_{il} + \varepsilon} \quad (4)$$

$$i = 1, \dots, N, j = 1, \dots, M$$

式中： $\hat{\mathbf{B}}$ 为稀疏化后的中间矩阵； ε 为防止分母为零的微小常数。

为保持每个位置上的注意力权重仍为概率分布形式，需要对 $\hat{\mathbf{B}}$ 的每一行重新归一化，得到稀疏注意力矩阵 $\mathbf{B}^{(s)}$ 。该稀疏锚注意力一方面在表示空间中引入结构化稀疏性，使每个时间-频率位置仅依赖少数关键锚点，提升潜在表示对异常模式的对比度；另一方面减弱无关背景模式对重构过程的干扰，使异常在注意力图和重构误差分布中呈现出更清晰的分离边界。

2.3 模型总体设计

模型总体构型如图2所示，FA-SAE以预先提取的声学特征向量 $\mathbf{x}$ 作为输入，首先经过两级 Linear+ReLU 映射到隐含表示空间。

随后进入滤波子网络：多条并行分支在时域与频域上执行卷积、频谱变换和逐通道映射，得到结构先验 $\mathbf{S}$ ，与原始支路输出 $\mathbf{Z}$ 逐点融合后，经一维卷积与残差连接形成经噪声抑制和谱形重整的特征。上层多头自注意力对不同时间-频率位置的相关性进行重加权，生成全局编码序列。解码阶段通过稀疏锚注意力将编码序列投影到少量锚向量，结合 Top-k 锚筛选仅保留主要关联，再经前馈网络、归一化层及线性投影逐步恢复到原始维度，得到重构向量。最终，以输入与重构之间的均方误差作为异常度量：

$$e(\mathbf{x}) = \frac{1}{D} \sum_{d=1}^D (\mathbf{x}_d - \hat{\mathbf{x}}_d)^2 \quad (5)$$

式中： $\mathbf{x} \in \mathbb{R}^D$ 为输入特征向量； $\hat{\mathbf{x}}$ 为模型输出； D 为特征维度。

3 实验与分析

3.1 数据集

本文选用公开工业设备异常声音数据集 MIMII 中的滑轨子集 6 dB 场景作为研究对象。具体信息如表1所示。本研究在数据划分上采用典型的无监督异常检测设定，即仅使用正

常样本进行模型训练，将异常样本严格保留用于测试。在每个包含正常与异常子目录的子数据集内，首先按异常样本数选取等量正常样本，与全部异常样本共同构成标签均衡的评估集，其余正常样本全部划入训练集。

表1 数据集信息

设备 ID	正常数量	异常数量
id_00	1 068	356
id_02	1 068	267
id_04	534	178
id_06	534	89

3.2 实验环境

本研究在一台配备 NVIDIA A10 专用计算加速卡的 Linux 服务器（ubuntu22.04）上完成实验。运行环境采用 Python3.11。深度学习框架使用 PyTorch2.8.0，CUDA12.6，并配套 torchaudio2.8.0 与 torchvision0.23.0。

3.3 评估指标

本研究采用接收者工作特征曲线下面积（area under the roc curve, AUC）及其在低误报区间上的局部面积（partial AUC, pAUC）作为主要评估指标，用公式表示为：

$$\text{AUC} = \frac{1}{N^- N^+} \sum_{i=1}^{N^-} \sum_{j=1}^{N^+} H(\mathbf{A}_\theta(\mathbf{x}_j^+) - \mathbf{A}_\theta(\mathbf{x}_i^-)) \quad (6)$$

$$\text{pAUC} = \frac{1}{|pN^-| N^+} \sum_{i=1}^{|pN^-|} \sum_{j=1}^{N^+} H(\mathbf{A}_\theta(\mathbf{x}_j^+) - \mathbf{A}_\theta(\mathbf{x}_i^-)) \quad (7)$$

3.4 实验结果

为验证本文所提算法的有效性，本文将其与多种典型自编码器模型进行了对比，包括基本自编码器（AE）、深层自编码器（DEEP_AE）、时序建模的 LSTM_AE 以及引入残差结构的 RESIDUAL_AE。同实验设置下，本文模型取得 96.51% 的 AUC 和 89.10% 的 pAUC，相比最基础 AE 分别提升 9.25 和 13.11 个百分点，相比当前最优 RESIDUAL_AE 亦有 1.55 和 5.06 个百分点的增益，尤其在低假阳性率区域的 pAUC 上优势更为显著。对比实验结果如表2所示。

表2 对比实验

模型	AUC/%	pAUC/%
AE	87.26	75.99
DEEP_AE	90.66	79.20
LSTM_AE	93.09	81.39
RESIDUAL_AE	94.96	84.04
本文模型	96.51	89.10

本文围绕 Attention-AE 基线模型依次引入 PairFilter 与（稀疏）锚注意力结构，对各组件的贡献进行了系统消融，如表 3 所示，其中 PF 表示 PairFilter；Anch 表示 Anchor Attention 模块；SA 表示 Sparse Anchor Attention 模块。Baseline(attention-ae) 的 AUC/pAUC 为 92.90%/79.41%。首先，仅加入 PairFilter 时，AUC 提升至 94.56%，pAUC 明显升至 85.66%，表明多分支滤波与门控残差能够有效抑制背景扰动、提升低假阳性率区间的判别能力。仅加入 Anchor Attention 后，AUC 与 pAUC 分别达到 95.90% 与 88.02%，说明显式建模少量“锚”关联可以更充分地捕获正常声音模式的关键结构，从而对异常偏离更加敏感。在此基础上，将锚注意力稀疏化的 Sparse Anchor Attention 进一步将 AUC 提升到 96.34%、pAUC 提升到 88.21%，验证了 Top-k 稀疏筛选在抑制冗余注意力扩散并突出关键依赖关系方面的有效性。将 PairFilter 与锚注意力联合使用时，AUC/pAUC 为 94.95%/86.02%，整体优于单独使用 PairFilter，体现出频域滤波与结构注意力之间的互补性；而“PairFilter+Sparse Anchor Attention”组合获得最高的 96.51%/89.10%，说明在统一框架下同时进行频谱整形与稀疏结构建模，可以在保持整体区分度的同时，显著提升实际应用中更为关键的低 FPR 区间检测鲁棒性。

表 3 消融实验

模型配置	AUC/%	pAUC/%
Baseline(attention-ae)	92.90	79.41
+PF	94.56	85.66
+Anch	95.90	88.02
+SA	96.34	88.21
+PF+Anch	94.95	86.02
+PF+SA	96.51	89.10

4 结语

本文提出了一种结合 PairFilter 和稀疏锚注意力机制的自编码器音频异常检测方法，面向 MIMII 数据集中的滑轨设备，在无监督框架下实现对低噪声工况下异常声纹的检测。通过引入 PairFilter 模块，对输入 Mel 频谱进行频域重整与噪声抑制，提升模型对关键结构成分的响应能力，抑制背景中周期性机械噪声与无关细节的干扰，从而降低误检率；采用稀疏锚注意力对潜在空间中的少量关键“锚”关系进行建模与 Top-k 筛选，将正常运行模式的结构先验显式编码，使模型能够有效分离正常与异常声学模式，增强对细微异常的检测能力。未来将继续在复杂噪声与跨设备场景下优化网络结构与训练策略，进一步轻量化模型并与健康指数可视化软件系统相融合。

参考文献：

- [1] 肖遥, 彭焘, 冯时, 等. 基于对比学习的自监督机器异常声音检测[J]. 复旦学报(自然科学版), 2025, 64(5): 579-590.
- [2] PUROHIT H, TANABE R, ICHIGE K, et al. MIMII dataset: sound dataset for malfunctioning industrial machine investigation and inspection[EB/OL]. (2019-09-20)[2025-12-30]. <https://doi.org/10.48550/arXiv.1909.09347>.
- [3] ZAMAN K, LI K, SAH M, et al. Transformers and audio detection tasks: an overview[J]. Digital signal processing, 2025, 158:104956.
- [4] WANG M, MEI Q S, SONG X Y, et al. A machine anomalous sound detection method using the IMS spectrogram and ES-MobileNetV3 network[J]. Applied sciences-basel, 2023, 13(23): 12912.
- [5] 王亚威, 张巧灵. 基于特征融合和流模型的无监督设备异常声音检测[J]. 电子科技, 2025, 38(11): 87-95.
- [6] 史爱武, 马淑然. 基于多特征融合与插值卷积自编码器的机械异常声音检测研究[J]. 软件导刊, 2025, 24(2): 40-47.
- [7] GUAN J, LIU Y D, KONG Q Q, et al. Transformer-based autoencoder with ID constraint for unsupervised anomalous sound detection[J/OL]. EURASIP journal on audio, speech, and music processing, 2023[2025-12-23]. <https://link.springer.com/article/10.1186/s13636-023-00308-4>.
- [8] YI K, FEI J R, ZHANG Q, et al. FilterNet: harnessing frequency filters for time series forecasting[EB/OL]. (2024-11-05)[2025-11-23]. <https://doi.org/10.48550/arXiv.2411.01623>.
- [9] SHAN J Q, WANG J X, ZHAO L F, et al. AnchorFormer: differentiable anchor attention for efficient vision transformer[J]. Pattern recognition letters, 2025, 197: 124-131.
- [10] CHEN X, LI H, LI M Q, et al. Learning a sparse transformer network for effective image deraining[EB/OL]. (2023-03-21)[2025-12-30]. <https://doi.org/10.48550/arXiv.2303.11950>.
- [11] JAYAKUMAR S M, PASCANU R, RAE J W, et al. Top-KAST: Top-K always sparse training[EB/OL]. (2021-06-07)[2025-12-30]. <https://doi.org/10.48550/arXiv.2106.03517>.

【作者简介】

彭瑜(2000—), 男, 安徽泗县人, 硕士研究生, 研究方向: 信号处理、异常检测。

刘豪睿(1978—), 男, 山东德州人, 博士, 副教授、硕士生导师, 研究方向: 设备运行监测、人工智能。

(收稿日期: 2025-11-21 修回日期: 2026-01-12)