Computer Science and Application 计算机科学与应用, 2023, 13(11), 2089-2096
Published Online November 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1311208

Hans汉斯

基于深度学习的机器异常声音检测

朱　鹏，黎春玲，郑荣璞，刘　琳*，魏喜庆，吕　品

上海电机学院电子信息学院，上海

收稿日期：2023年10月17日；录用日期：2023年11月16日；发布日期：2023年11月23日

摘　要

随着大规模工业生产的发展，机器设备的健康管理越来越重要。由于机器设备潜在的故障，机器异常声音的检测对工业生产的保障有待提高。不同的机器运作时发出的声音有规律性，可以根据这一特性判断机器是否处于一个正常运作状态，通过对机器运作时的声音特征进行研究，提出一种基于深度学习的机器异常声音的检测，通过对声音特征的提取，经过模型的训练，判断机器是否处于异常状态，防患于未然。首先对数据集通过等高梅尔滤波器处理后提取出对数Mel谱作为声音特征，之后针对实际中异常声音的缺失等问题，使用mobilenetv2对声音模型进行训练，通过模型输出的逻辑回归值来计算异常分数和确定异常阈值。经过对比分析，表明对原始音频进行特征提取后训练的模型，机器异常声音检测性能有所提升。

关键词

异常声音检测，深度学习，对数Mel谱，Mobilenetv2

A Machine Abnormal Sound Detection Based on Deep Learning

Peng Zhu, Chunling Li, Rongpu Zheng, Lin Liu*, Xiqing Wei, Pin Lv

School of Electronic Information Engineering, Shanghai Dianji University, Shanghai

Received: Oct. 17th, 2023; accepted: Nov. 16th, 2023; published: Nov. 23rd, 2023

Abstract

With the development of large-scale industrial production, the health management of machinery and equipment is becoming increasingly important. Due to the potential failure of machinery and equipment, the detection of abnormal machine sounds needs to be improved. The sound emitted

*通讯作者。

文章引用: 朱鹏, 黎春玲, 郑荣璞, 刘琳, 魏喜庆, 吕品. 基于深度学习的机器异常声音检测[J]. 计算机科学与应用, 2023, 13(11): 2089-2096. DOI: 10.12677/csa.2023.1311208

by different machines during operation is regular, and whether the machine is in a normal operation state can be judged according to this characteristic. Through the research on the sound characteristics of the machine during operation, a machine abnormal sound detection based on deep learning is proposed. Through the extraction of sound characteristics and the training of the model, whether the machine is in an abnormal state can be judged to prevent potential problems. Firstly, the data set is processed by the constant-height Mel filter to extract the logarithmic Mel spectrum as the sound feature, then the sound model is trained by mobilenetv2 to process the absence of abnormal sound in the data set, and the abnormal score and the abnormal threshold are calculated by the logistic regression value output by the model. After comparative analysis, it is shown that the machine abnormal sound detection performance of the model trained is improved by feature extraction of the original audio.

Keywords

Abnormal Sound Detection, Deep Learning, Log-Mel Spectrum, Mobilenetv2

1. 绪论

1.1. 研究背景

随着现代化水平的迅速提高，安全监控系统在生活中的应用越来越普遍。目前，监控系统主要依赖于视频信息来进行监控，但是基于视频的系统存在一些问题。相比图像信号，声音信号的计算量较小，包含更多信息，并且更具隐私性。此外，音频设备相对简单且稳定性较高，异常声音能够有效地揭示异常情况和突发事件，从而弥补视频监控的不足。音频监控系统通过检测和识别区域内的声音信息来进行监控，实现更高的自动化和智能化，对于智能监控系统和安全防护至关重要。近年来，工业机器在国防建设、石油化工、电力部门以及日常生活的交通出行中得到了广泛应用。各种工业机器不断朝着集成化、大型化和一体化的方向发展，它们的结构变得越来越复杂。一旦出现故障，不仅设备会失效，还可能导致巨大的经济损失和严重的安全事故。声音异常检测可以及时准确地判断机器是否发生故障，提前预测和处理潜在的故障状态，预防性地维护可能出现的故障，从而减少维修成本和支出。此外，通过实时监测和定期检查机器状态，可以在不影响整体生产的情况下进行维修，防止故障进一步恶化，延长机械设备的使用寿命，最大化使用效率。声音异常检测还能了解机器运行状态的变化，确保机器连续稳定地工作，有助于更精确地管理和控制机器。

1.2. 研究现状

为了可以通过提取机器运行的声音特征来判断机器是否出现问题。在最近的几年的研究中，从模型角度出发对声音识别的研究可以分为 3 类：① 将其底层声学特征输入到通用背类模型 GMM-UBM [1] [2] 中，建立对应语种的识别模型以进行识别[3]。② 基于端到端的语种识别模型，提取深度语种特征并进行识别。③将声学特征图像化，借助卷积神经网络进行语种识别库设置和帧移动下的语种识别[4]。从特征角度出发对语种识别进行研究的学者则使用梅尔频率倒谱系数(MFCC, Me-scale frequency cepstral coefficients) [5]，伽马频率倒谱系数(GFCC, Gammatone frequency cepstral coefficients) [6]等单一声学特征进行语

种识别。但在这些年的发展中，研究者们也不仅仅将目光局限在单一的方法之中，从而出现了很多具有融合特征的方法[7]-[13]，例如孙颖团队将 MFCC 和韵律的特征融合[7]；张科团队将 MFCC 和 GFCC 特征融合[8]；MTS 等将 MFCC 和能量归一化倒诺系数(PNCC, power normalized cepstral coefficient)融合[9]，从而进行语音识别。但以上融合特征的方法仍然有着诸多缺陷，比如这些方法都是利用提升特征的维度但牺牲了算力来提升语种识别性能。所以也有很多研究者开始转向如何在融合特征的基础上并且降维提高高噪声下的音频识别速率和准确率[14] [15] [16]。如周萍，沈昊，郑凯鹏的基于 MFCC 与 GFCC 混合特征参数的说话人识别[14]，Anirban，Bhowmick 的基于奇异值分解的特征嵌入的印度区域语言的识别[15]，邵玉斌，刘晶，龙华，等的面向实噪声环境的语种识别[16]。

2. 相关理论和技术

2.1. 特征提取

MFSC (log Mel-frequency spectral coefficients)的提取过程包括预处理、快速傅里叶变换、Mel 滤波器组、对数运算、动态特征提取共五个步骤：

① 预处理过程：是对其原始音频数据进行数字化、预滤波、预加重、端点检测、分帧、加窗等操作，使其信号特征更加明显，去除冗余数据。

$$H(z)=1-\mu z^{-1} \tag{1}$$

② 快速傅里叶变换过程：快速傅里叶变换即利用计算机计算离散傅里叶变换(DFT)的高效、快速计算方法的统称，简称 FFT。将音频从时域转换为频域。

$$f(\omega)=f\left[f(t)\right]=\int_{\infty}^{\infty}f(t)\mathrm{e}^{-i\omega t}\mathrm{d}t \tag{2}$$

③ Mel 滤波器组过程：研究表明，人类对频率的感知并不是线性的，并且对低频信号的感知要比高频信号敏感。对 1 kHz 以下，与频率成线性关系，对 1 kHz 以上，与频率成对数关系。频率越高，感知能力就越差。为了模拟人耳的听觉机制。从而研制出来了 Mel 滤波器组。所以，Mel 滤波器组的在低频密集，高频稀疏。

$$Mel(f)=2595\times\lg\left(1+\frac{f}{700}\right) \tag{3}$$

④ 对数运算：对数运算包括取模和 log 运算。将原始语音信号经过傅里叶变换得到频谱：取模是仅使用幅度值，忽略相位的影响，因为相位信息在语音识别中作用不大。log 运算是为了分别包络和细节，包络代表音色，细节代表音高。显然语音识别是为了识别音色。另外，人的感知与频域的对数成正比，正好使用 log 运算对上述梅尔语谱图的纵轴进行对数缩放，可以放大低频率处的能量差异。

⑤ 动态特征提取：标准的倒谱参数 MFSC 只反映了语音参数的静态特性，语音的动态特性可以用这些静态特征的差分谱来描述。实验证明：把动、静态特征结合起来才能有效提高系统的识别性能。差分参数的计算可以采用下面的公式：

$$d_t=\begin{cases}C_{t+1}-C_t, t<K\\ \dfrac{\sum_{K=1}^{K}K\left(C_{t+k}-C_{t-k}\right)}{\sqrt{2\sum_{K=1}^{K}k^2}}, \text{其他}\\ C_t-C_t, t\geq Q-\mathrm{K}\end{cases} \tag{4}$$

其中，d_t 表示第 t 个一阶差分，C_t 表示第 t 个倒谱系数，Q 表示倒谱系数的阶数，K 表示一阶导数的时间

差，可取 1 或 2。将上式的结果再代入就可以得到二阶差分的参数。

因此，MFSC 的全部组成其实是由：N 维 MFSC 参数(N/3 个 MFSC 系数 ＋N/3 个一阶差分参数 ＋N/3 个二阶差分参数) ＋ 帧能量(此项可根据需求替换)。这里的帧能量是指一帧的音量(即能量)，也是语音的重要特征。

2.2. 深度学习模型

2.2.1. 自编码器(AE)

自编码器(Autoencoder, AE)是一个无监督学习模型，通过对大量的无标签数据压缩重构，自动学习数据特征，因此该神经网络模型在机器异常检测领域中被广泛使用。该模型由输入层、隐藏层和输出层组成，通过使用神经网络对数据进行编码和解码[17]。自编码器结构如图 1 所示。隐藏层输出作为样本的抽象特征表示，AE 首先通过输入层接收样本，隐藏层将其转换成高效的抽象表示，而后输出层将原始样本重构并对输入进行特征学习，并将学习到的特征重组作为输出，目标是降低输入和重构输出之间的损失，使重构的输出与输入的相似度越来越接近[18]。

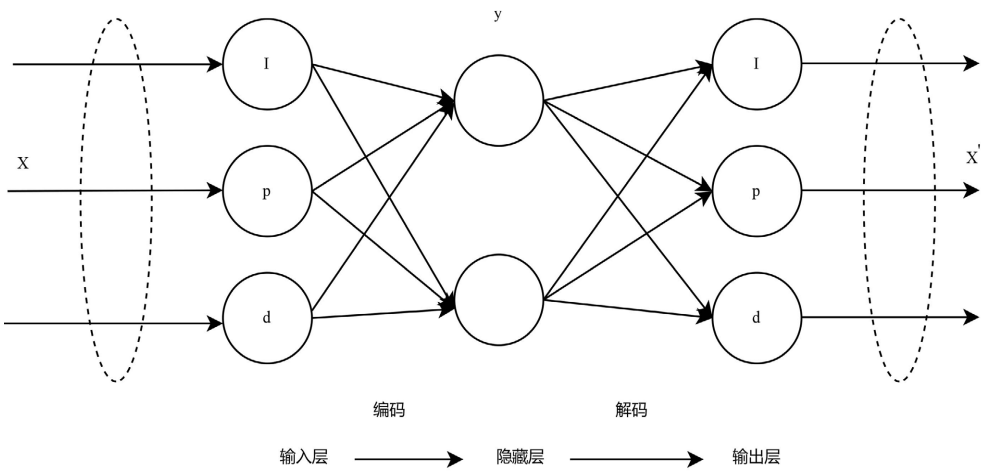

Figure 1. Autoencoder structure
图 1. 自编码器结构

自编码的过程可以分为两部分：输入层到隐藏层为编码过程和隐藏层到输出层为解码过程。编码过程是将高维的输入样本映射到低维，以此实现样本压缩与降维；解码过程则是将抽象表示样本转换为期望输出，目的是输入样本的重构。

在训练阶段中，编码过程对输入样本进行编码操作得到编码层；解码过程对编码层进行解码操作得到输入样本的重构，并通过调整网络参数使重构误差达到最小值，获得输入特征的最优抽象表示。

编码过程如下：

$$y = f\left(wx + b\right) \tag{5}$$

解码过程如下：

$$x' = f\left(w'y + b'\right) \tag{6}$$

式中：x 表示 AE 的输入；y 表示隐藏层；x' 表示重构数据；$f\left(x\right)$ 表示非线性激活函数；w、w'、b、b' 表示神经网络参数。在重构过程中，若要使重构输出的 x' 和输入 x 尽可能一致，需要用最小化负对数

似然的损失函数来训练模型。

$$L = -\log P\left(x \middle| x'\right) \tag{7}$$

2.2.2. MobileNetV2

MobileNet 网络模型在 2017 年被谷歌团队提出，该模型是一个轻量级深度神经网络模型，具有精度高、计算量小、体积小，适用于运算能力有限和资源受限的平台，目前已有的三个版本，分别为MobileNetV1、MobileNetV2、MobileNetV3 [17]。

MobileNetV2 由谷歌团队在 2018 年提出的一种新型卷积神经网络，可用于目标检测和分割，相比于传统的深度神经网络，该模型准确率更高，推理速度更快和模型更小。MobileNetV2 在 MobileNetV1 的基本概念构建上，使用在深度上可分离的卷积作为高效的构建块，基于大量的反向残差结构对数据特征进行提取，设计了反向残差块(Inverted resi-dual block)和线性瓶颈(Linear bottleneck)，提高了检测的精度。反向残差块结构如图 2 所示。因为反向残差模块是中间维度大，两边维度小，反向残差块将低维特征使用点卷积 1 × 1 升维，之后使用深度卷积 3 × 3 + Relu 对信息特征进行提取，最后使用点卷积 1 × 1 + Linear 对特征再降维，得到本层特征的输出，并进行一次点卷积的结果和输入相加。

线性瓶颈层的输入通过点卷积 1 × 1 + Relu 进行升维，从 k 维增加到 k'' 维；之后通过点卷积 3 × 3 + Relu 可分离卷积对样本进行特征提取采样(stride > 1 时)，此时特征维度已经为 k'' 维度，最后通过点卷积 1 × 1 (无 Relu)进行降维，维度从 k'' 降低到 k' 维。为防止使用 Relu 的某通道的信息进行处理后会不可避免的有信息丢失，为此在最后降维中使用 Linear 替代 Relu，以防止 Relu 破坏了数据的特征，使有效信息得到保留[19]。

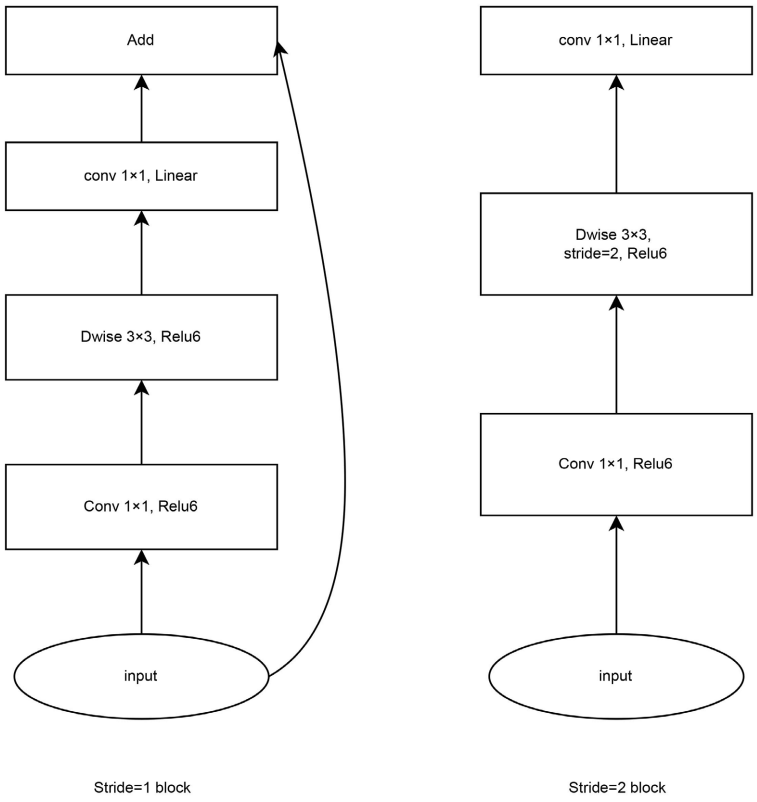

Figure 2. Iverse residual block structure
图 2. 反向残差块结构

3. 实验与设置

梅尔标度滤波器由多个三角滤波器组成滤波器组，低频处滤波器密集，门限值大，高频处滤波器稀疏，门限值低。恰好对应了频率越高人耳越迟钝这一客观规律。正常所用的滤波器形式叫做等面积梅尔滤波器(Mel-filter bank with same bank area)，在人声领域(语音识别，说话人辨认)等领域应用广泛，但是如果用到非人声领域，就会丢掉很多高频信息。因而本文采用等高梅尔滤波器(Mel-filter bank with same bank height)进行实验。

3.1. 实验数据集

数据集使用 DCASE Challenge 2021 Task2 开发数据集，由 MIMII 数据集和 ToyADMOS 数据集组成。数据集包括开发数据集和评估数据集。MIMII 数据集是用于故障工业机器调查和检查的声音数据集，记录了 7 种不同类型的工业机器的正常和异常声音，即 fan、gearbox、pump、slider、valve 和 ToyADMOS 数据集的两种机器类型(即 ToyCar 和 ToyTrain)。每种数据中包含完训练集和测试数据集，即 train、source_test、target_test，每个数据集中又设置 3 个类别的声音 Section00、Section01、Section02，每段声音约 10 秒，包含了机器运作声音和环境噪声[20] [21] [22] [23]。

3.2. 参数设置

采用亚当(Adam)优化算法，学习率为 0.00001。训练时迭代次数设置为 20 次，每一批数据的大小(Batch size)为 32。

3.3. 评估指标

本实验使用 AUC 和 pAUC 进行评估。pAUC 是根据预先指定目标范围内的 ROC 曲线的一部分计算得出的 AUC。在本次实验的指标中，pAUC 计算为低假阳性率(FPR)范围内的 AUC [0, p]，p 为 0.1。每种机器类型、类别和域的 AUC 和 pAUC 定义为

$$AUC(m,n,d)=\frac{1}{N_{-}N_{+}}\sum_{i=1}^{N_{-}}\sum_{j=1}^{N_{+}}H\left(A\theta\left(x_j^{+}\right)-A\theta\left(x_i^{-}\right)\right) \tag{8}$$

$$pAUC(m,n,d)=\frac{1}{\lfloor pN_{-}\rfloor N_{+}}\sum_{i=1}^{pN_{-}}\sum_{j=1}^{N_{+}}H\left(A\theta\left(x_j^{+}\right)-A\theta\left(x_i^{-}\right)\right) \tag{9}$$

m 表示机器类型；n 表示机器的某个类别；$d=$ {源域，目标域}用于确定一个域；$H(x)$当 $x>0$ 时返回 1，否则返 0；$\left\{x_i^{-}\right\}_{i=1}^{N_{-}}$ 表示某个机器类型某个类别某个域的正常样本，$\left\{x_j^{+}\right\}_{j=1}^{N_{+}}$ 表示某个机器类型某个类别某个域的异常样本；N_{-} 表示某个机器类型某个类别某个域的正常样本数量，N_{+} 表示某个机器类型某个类别某个域的异常样本数量。

3.4. 结果分析

Table 1. Comparison of experimental results of some machines with baseline systems
表 1. 部分机器的实验结果与基线系统的对比

机器类型	ToyCar (AE)	Valve (AE)	ToyCar (MobileNetV2)	Valve (MobileNetV2)	ToyCar (本文方法)	Valve (本文方法)
AUC	63.19%	53.74%	59.58%	57.07%	63.19%	57.05%
pAUC	52.42%	50.61%	57.64%	52.83%	58.54%	56.04%

本文所采用的方法通过对比实验，从表 1 的结果可看出 ToyCar、Valve 的算数平均 pAUC 分别提升了 0.9%，3.21%。

4. 总结

本作品描述了一种基于深度学习的机器异常声音检测方法，主要是对音频数据集进行声音特征提取后使用自编码器(AE)模型和 MobileNetV2 模型进行训练，将训练好的模型使用开发数据集的测试数据进行评估检测模型。再通过模型输出的逻辑回归值来计算异常分数和确定异常阈值，然后判断机器是否异常。实验证明了通过对原始音频进行特征提取后训练的模型可以获得一定的异常检测效果。在未来的工作中，将考虑到数据集容量，需要扩大数据量，进一步检验方法的有效性，同时考虑采用更加合理的网络来提升模型的整体性能，进一步发挥深度学习网络的优势，并选择更加适合的分类器，以进一步提高异常检测的效果。

基金项目

上海大学生创新创业训练项目(项目编号：S202211458045)，上海电机学院电子信息学院“工业大数据开发技术”课程建设。

参考文献

[1] Wondimu, M. and Tekeba, M. (2019) Signal Based Ethiopian Languages Identification Using Gaussian Mixture Model. *Zede Journal*, **37**, 39-54.

[2] 邵玉斌, 刘晶, 龙华, 等. 基于声道频谱参数的语种识别[J]. 北京邮电大学学报, 2021, 44(3): 112-119.

[3] Jiang, B., Song, Y., Wei, S., *et al*. (2014) Deep Bottleneck Features for Spoken Language Identification. *PLOS ONE*, **9**, e100795. https://doi.org/10.1371/journal.pone.0100795

[4] Das, H.S. and Roy, P. (2021) A CNN-BiLSTM Based Hybrid Model for Indian Language Identification. *Applied Acoustics*, **182**, Article ID: 108274. https://doi.org/10.1016/j.apacoust.2021.108274

[5] Koolagudi, S.G. (2012) Identification of Language Using Mel-Frequency Cepstral Coefficients (MFCC). *Procedia Engineering*, **38**, 3391-3398. https://doi.org/10.1016/j.proeng.2012.06.392

[6] 张卫强, 刘加. 基于听感知特征的语种识别[J]. 清华人学学报(自然科学版), 2009, 49(1): 78-81.

[7] 孙颖, 姚慧, 张雪英, 等. 基于混沌特性的情感语音特征提取[J]. 天津大学学报(自然科学与工程技术版), 2015, 48(8): 681-685.

[8] 张科, 苏雨, 王靖宇, 等. 基于融合特征以及卷积神经网络的环境声音分类系统研究[J]. 西北工业大学学报, 2020, 38(1): 162-169.

[9] Al-Kaltakchi, M.T.S., Woo, W.L., Dlay, S.S., *et al*. (2016) Study of Fusion Strategies and Exploiting the Combination of MFCC and PNCC Features for Robust Biometric Speaker Identification. 2016 4*th International Conference on Biometrics and Forensics* (*IWBF*), Limassol, 3-4 March 2016, 1-6. https://doi.org/10.1109/IWBF.2016.7449685

[10] 郑艳, 姜源样. 基于特征融合的说话人聚类算法[J]. 东北大学学报(自然科学版), 2021, 42(7): 952-959.

[11] Bhanja, C.C., Bisharad, D., Laskar, R.H., *et al*. (2019) Deep Residual Networks for Pre-Classification Based Indian Language Identification. *Journal of Intelligent & Fuzzy Systems*: *Applications in Engineering and Technology*, **36**, 2207-2218. https://doi.org/10.3233/JIFS-169932

[12] Montavon, G. (2009) Deep Learning for Spoken Language Identification. *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, Vancouver, December 2009, 1-4.

[13] Deepti, D. (2020) A Language Identification System Using Hybrid Features and Back-Propagation Neural Network. *Applied Acoustics*, **164**, Article ID: 107289. https://doi.org/10.1016/j.apacoust.2020.107289

[14] 周萍, 沈昊, 郑凯鹏. 基于 MFCC 与 GFCC 混合特征参数的说话人识别[J]. 应用科学学报, 2019, 37(1): 24-32.

[15] Anirban, B. (2021) Identification/Segmentation of Indian Regional Languages with Singular Value Decomposition Based Feature Embedding. *Applied Acoustics*, **176**, Article ID: 107864. https://doi.org/10.1016/j.apacoust.2020.107864

[16] 邵玉斌, 刘晶, 龙华, 等. 面向实噪声环境的语种识别[J]. 北京邮电大学学报, 2021, 44(6): 134-140.

[17] 柯凯航. 基于深度学习的机器异常声音检测研究[D]: [硕士学位论文]. 汕头: 汕头大学, 2022. https://doi.org/10.27295/d.cnki.gstou.2022.000786

[18] 薛英杰, 陈颀, 周松斌, 等. 基于自监督特征提取的机械异常声音检测[J]. 激光与光电子学进展, 2022, 59(12): 361-371.

[19] Sandler, M., Howard, A., Zhu, M., *et al*. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[20] Purohit, H., Tanabe, R., Ichige, T., *et al*. (2019) MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection. https://doi.org/10.33682/m76f-d618

[21] Tanabe, R., Purohit, H., Dohi, K., Endo, T., Nikaido, Y., Nakamura, T. and Kawaguchi, Y. (2021) MIMII DUE: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection with Domain Shifts Due to Changes in Operational and Environmental Conditions. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), New Paltz, 17-20 October 2021, 21-25. https://doi.org/10.1109/WASPAA52581.2021.9632802

[22] Harada, N., Niizumi, D., Takeuchi, D., Ohishi, Y., Yasuda, M. and Saito, S. (2021) ToyADMOS2: Another Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection under Domain Shift Conditions. *Proceedings of the* 6*th Detection and Classification of Acoustic Scenes and Events* 2021 *Workshop* (*DCASE*2021), Barcelona, November 2021, 1-5.

[23] Kawaguchi, Y., Imoto, K., Koizumi, Y., Harada, N., Niizumi, D., Dohi, K., Tanabe, R., Purohit, H. and Endo, T. (2021) Description and Discussion on DCASE 2021 Challenge Task 2: Unsupervised Anomalous Detection for Machine Condition Monitoring under Domain Shifted Conditions. *Proceedings of the* 6*th Detection and Classification of Acoustic Scenes and Events* 2021 *Workshop* (*DCASE*2021), Barcelona, November 2021, 186-190.