Data-driven Non-uniform Filterbanks Based on F-ratio for Machine Anomalous Sound Detection

Kai Li[1], Dung Kim Tran[1], Xugang Lu[2], Masato Akagi[1], and Masashi Unoki[1*]
[1]*School of Information Science, Japan Advanced Institute of Science and Technology, Japan*
[2]*Advanced Speech Technology Laboratory, National Institute of Information and Communications Technology, Japan*
[1]{*kai_li, kimdungtran, akagi, unoki*}*@jaist.ac.jp* [2] xugang.lu@nict.go.jp

***Abstract*—Anomalous sound detection (ASD) aims to detect unknown anomalous sounds emitted from a target machine. Most advanced ASD systems use a complicated neural-network-based detector with the log Mel spectrum as input. However, different types of machines have different vibration frequency regions depending on their physical property. The Mel filterbank (FB), which has high resolution in low-frequency regions and low resolution in high frequency, may filter out discriminative information from some important frequency regions, particularly the high-frequency regions. We propose to quantify the frequency importance in ASD of seven types of machines using the Fisher's ratio (F-ratio). The quantified frequency importance is then used to design an ensemble of machine-wise non-uniform FBs and extract the log non-uniform spectrum (LNS). This LNS feature is input to an autoencoder NN-based detector for anomalous sound detection. Experimental results in the DCASE2022 Challenge Task 2 verify the correctness of the quantification results and the effectiveness of the proposed LNS. With a simple autoencoder-based detector, the performance in the averaged harmonic mean of the area under the ROC curve achieved a relative improvement of 9.22 and 5.60% in development and evaluation datasets, respectively.**

Index Terms: Anomalous sound detection, frequencies importance, data-driven non-uniform filterbank, log non-uniform spectrum, F-ratio.

I. Introduction

Anomalous sound detection (ASD) for machine condition monitoring enables workers to arrange maintenance work to fix machine problems in the earliest stages of anomaly, thus reducing maintenance costs and preventing damage. Developing advanced ASD systems is an important component of the fourth industrial revolution and has received increasing attention in recent years [1], [2]. ASD can be classified into two types of problems [3], i.e., supervised ASD in which recordings of anomalous events to be detected are available in training, and unsupervised ASD in which recordings of the anomalous events are not available in training.

Most methods for ASD are based on an unsupervised autoencoder (AE) model [4]–[6] because of difficulties in collecting anomalous sounds that can cover all possible types of anomalies [7]. These methods are used to detect "*unknown*" anomalous sounds that have not been observed using reconstruction errors. However, because the training procedure does not incorporate anomalous sounds, the effectiveness of such models may be limited if the learned features also fit with the anomalous sounds.

Recently, many sophisticated models are adapted and applied to further improve the effectiveness of back-end detectors [8], [9]. For example, the WaveNet architecture was used by Hayashi, et al. [10]. [11] proposed an ASD approach that utilizes a denoising autoencoder (AE) architecture with both feed-forward and Long Short-Term Memory (LSTM) units. The self-supervised approaches were used in [12], [13] to provide some additional information, i.e., machine type and machine identity (ID). [14] proposed a new AE architecture, named as Fully-Connected U-Net, to replace the conventional AE model. However, the performance of the neural network (NN)-based ASD methods depends significantly on the discrimination of acoustic front-ends.

The log Mel spectrum (LMS) is widely used as an acoustic front-end in an NN-based ASD system [4], [6], [15]. The Mel filterbank (FB) is designed on the basis of the pitch perception of the human ear. It has a higher resolution in the low-frequency regions and a lower resolution in the high-frequency regions [16]. However, it can be argued that the human ear is not the most effective in detecting machine anomalies. Moreover, different types of machines have different vibration frequency regions depending on their physical property. Consequently, the discriminative information of sounds emitted from different types of machines may be encoded non-uniformly in the frequency domain. The Mel FB may filter out important information at the high-frequency regions, decreasing the performance of an ASD system. Therefore, quantifying the importance of the frequencies of different types of machines for ASD is necessary.

How can the importance of different frequencies be quantified? The Fisher's ratio (F-ratio) is a statistical-based method and widely used to measure the discriminative ability of a feature for pattern recognition [17]. It has been used to evaluate the importance of different frequencies in speaker recognition [18], [19], emotion recognition [20], and replay attack detection [21]. The calculation of the F-ratio requires no training data and is comparatively straightforward and efficient.

To extract more distinguished information from the frequency domain, we propose to quantify the frequency importance in ASD of sounds produced by seven types of machines using the F-ratio, called machine-wise F-ratio. We

*Corresponding author.

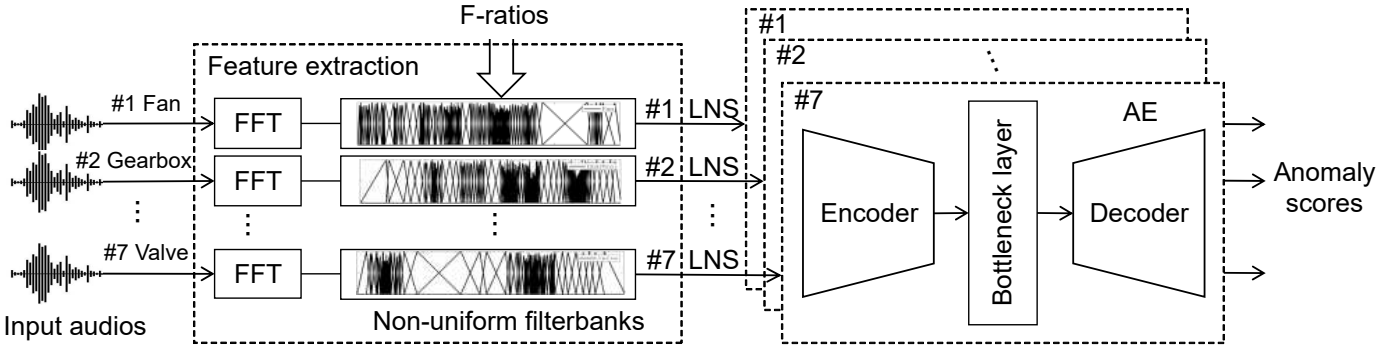

Fig. 1. Systems using LNSs extracted by the proposed data-driven non-uniform FBs with AE-based detectors for machine ASD.

are the first to quantify the frequency importance of machines using the data-driven method to detect anomalous sounds. With the quantification results, we aim to visualize where the discriminative features of each machine are encoded in the frequency domain. To highlight such important frequency bands, we design seven non-uniform FBs which have high resolutions in the frequency regions with high F-ratios and extract the log non-uniform spectrum (LNS). Experimental results in DCASE2022 Challenge Task 2 verify the correctness of quantification results and effectiveness of the proposed LNS.

II. Quantification of frequency-band importance using machine-wise F-ratio

The frequency bands with more discriminative features should possess high inter-class variances and low intra-class variances between normal and anomalous sound classes [18]. Therefore, we defined the F-ratio for machine m as

$$F_m = \frac{\frac{1}{2}\sum_c (u_{m,c} - u_m)^2}{\frac{1}{2N}\sum_c \sum_{i=1}^{N} (x_{m,c}^i - u_{m,c})^2}, \tag{1}$$

where $x_{m,c}^i$ is the sub-band energy of the i-th audio of class c with $i = 1, 2, ..., N$, $m \in \{\text{fan, gearbox, bearing, slider, toycar, toytrain, valve}\}$, and $c \in \{\text{normal, anomaly}\}$. The equations

$$u_{m,c} = \frac{1}{N}\sum_{i=1}^{N} x_{m,c}^i \qquad \text{and} \qquad u_m = \frac{1}{2N}\sum_c \sum_{i=1}^{N} x_{m,c}^i$$

are used to calculate the variables that represent the sub-band energy averages for class c and for all classes, respectively.

Equation (1) is the ratio between the inter-class variances and intra-class variances of speech power in a given frequency band. A larger value obtained in a frequency band means that more discriminative information is encoded in that band.

III. Design of data-driven non-uniform filterbanks for LNS extraction

To show the correctness of quantification results, we designed non-uniform FBs and used them to extract the LNS for each machine. The non-uniform FBs were designed by highlighting the frequency bands with relatively high F-ratios.

The distribution density of the triangular band-pass filters is assigned to be directly proportional to the F-ratios. The steps for designing a non-uniform FB are as follows: (1) calculate the weight k based on the F_m, $k = f_s/(2 \times \sum F_m)$, where f_s is the sampling frequency, (2) calculate the cumulative sum (CS) of the weighted F_m, CS = Cumsum$(k \times F_m)$, (3) fit the curve of the mapping frequency from the linear scale to the adaptive scale by using the cubic spline interpolation, (4) calculate the center frequencies of the triangular band-pass filters $C(j)$ on the basis of the fitting curve, and (5) design non-uniform FBs with the non-uniform resolutions.

Finally, the LNS was extracted by replacing the FB used in the extraction processes of the LMS. The detailed process of our proposed ASD systems are shown in Fig. 1, the AE model includes encoder, bottleneck layer, and decoder modules. All modules consist of fully connected layers. The mean squared error (MSE) is used as the cost function to optimize the overall system. In the testing phase, audio with high reconstruction error was treated as anomalous sound.

IV. Experimental setup

A. Datasets

We used the dataset provided by the DCASE2022 Challenge Task 2 [22], [23]. The dataset is comprised of normal and anomalous sounds produced by seven types of machines, i.e., fan, gearbox, bearing, slide, tor car, toy train, and valve. Sounds recorded from each type of machine are divided into six sections in accordance with the differences in machine configurations; sections 01, 02, and 03 are organized in the development dataset; sections 04, 05, and 06 are in the evaluation dataset. During the analysis step, we used {100 normal,100 anomaly}×3×7 clips of the development data to calculate F-ratios. During the training step, we used 1,000×3×7 clips of the development data to train the AE model. It is worth noting that the data used to calculate the F-ratio is not used for model training. During the evaluation step, we used 200×3×7 clips of test data in both the development and evaluation datasets to evaluate the effectiveness of the proposed method. The length of each clip was fixed to 10s.

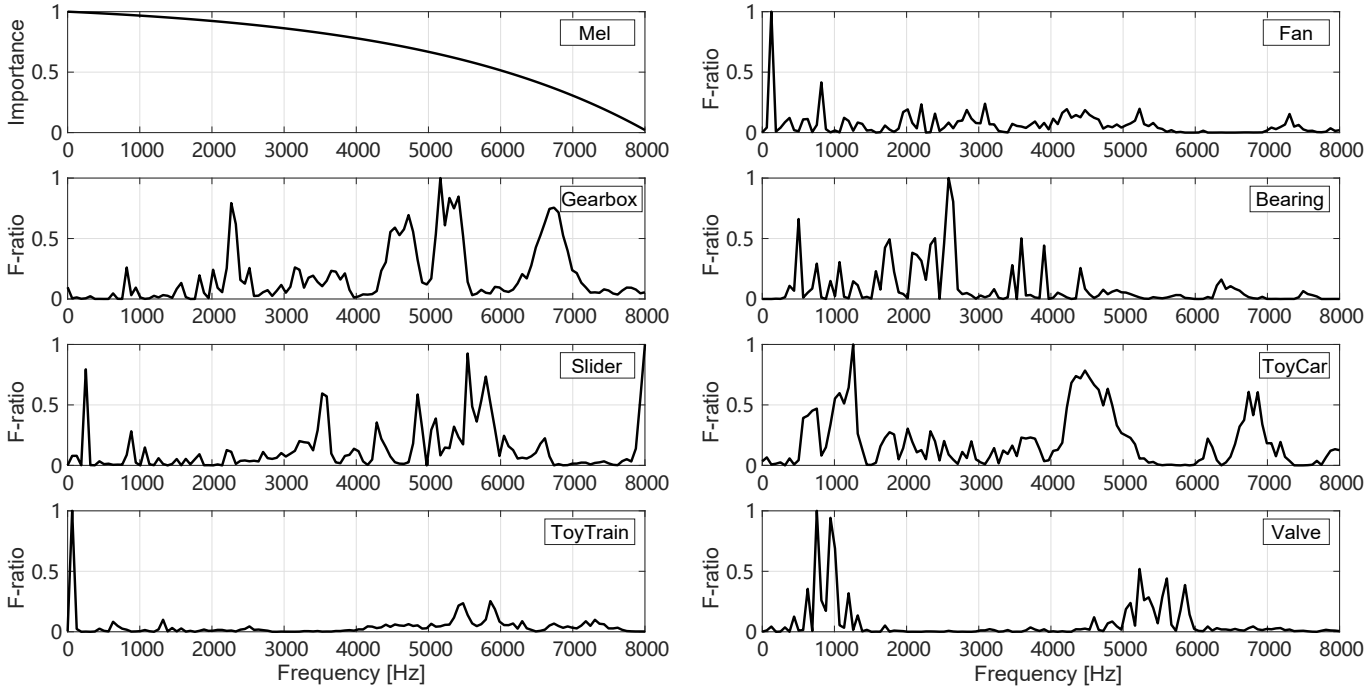

Fig. 2. Frequency-band importance of Mel scale and quantified frequency-band importance using machine-wise F-ratio for each machine. All frequency-band importances were normalized from 0 to 1.

B. Evaluation metrics

The area under the curve (AUC) and partial-AUC (pAUC) for the receiver operating characteristic (ROC) curves were used to evaluate the proposed ASD method. The formulae of AUC and pAUC can be found in [4]. Generally, AUC and pAUC are the average sums of anomaly scores. However, the difference between pAUC and AUC is that pAUC is designed to focus on a low false-positive-rate portion of the ROC curve over a pre-specified range of interest $[0, p = 0.1]$. In practical situations, if an ASD system generates false alarms frequently (high false-positive rate), the system is not trustworthy. Therefore, using pAUC to encourage a high true-positive rate under low false-positive-rate conditions is essential.

C. Experimental conditions

To extract the LMS, 10-s audio clips were first split into different frames with frame lengths of 64 ms and hop lengths of 32 ms. The Mel-spectrogram feature was then extracted with the following parameters: n_fft=1024, hop_length=512, $num\_filters$=128, and $power = 2.0$. We extracted the LNS using the same configuration but different FBs compared with the LMS. Five consecutive frames with a sliding window were concatenated into one feature vector with a dimension of 640 and fed into the detector. For example, we assume that the input signal is $X = \{X_t\}_{t=1}^{T}$ where $X_t \in \mathbb{R}^M$, and M and T are the number of Mel-filters and time-frames, respectively. Then, the acoustic feature at t is obtained by concatenating consecutive frames of the feature as $\psi_t \in \mathbb{R}^D$, where $D = P \times M$, $P = 5$, $M = 128$ and $D = 640$. The reconstruction error is calculated as:

$$E(X) = \frac{1}{DT} \sum_{t=1}^{T} \parallel \psi_t - r(\psi_t) \parallel_2^2, \tag{2}$$

where $r(\psi_t)$ is the vector reconstructed by the AE model, and $\| \cdot \|_2$ is $\mathcal{L}_2$ norm.

The AE model had four dense layers with 128 dimensions for the encoder/decoder and one bottleneck layer with 8 dimensions. We trained the model for 100 epochs using the Adam optimizer [24] with a learning rate of 0.0001 and batch size of 128.

V. Results

The quantification results of discriminative information for ASD for each machine are shown in Fig. 2. We also illustrate the frequency-band importance of the Mel scale for comparison. This result can be understood as the derivative of the Mel scale. It is evident that the frequency importance in the Mel scale decreased with the increase in frequency. The quantification results using the machine-wise F-ratio indicate that the discriminative feature for ASD of each type of machine was encoded non-uniformly in the frequency domain. There are many discriminative features concentrated in the high-frequency regions, such as the gearbox and slider.

Based on the F-ratios, we designed the machine-wise ASD systems on accordance with the pipeline shown in Fig. 1, and carried out experiments to examine its effectiveness. The results of harmonic mean (HM) and arithmetic mean (AM) in

TABLE I
OVERALL RESULTS BY USING THE PROPOSED (LNS) AND BASELINE (LMS) FEATURES IN TERMS OF AUC (%) AND PAUC (%) IN THE DEVELOPMENT DATASET.

Machines	Sections	AUC		pAUC	
		LMS	Proposed	LMS	Proposed
Toy car	AM	64.22	**66.48**	53.25	**56.11**
	HM	63.01	**65.00**	53.23	**56.05**
Toy train	AM	51.20	**57.35**	50.49	**52.65**
	HM	49.55	**56.74**	50.48	**52.62**
Bearing	AM	56.68	**65.75**	50.93	**58.18**
	HM	55.65	**65.12**	50.86	**57.70**
Fan	AM	**64.45**	60.68	**58.39**	56.79
	HM	**63.14**	59.23	**57.93**	56.40
Gearbox	AM	65.54	**70.10**	59.00	**59.93**
	HM	65.28	**69.79**	58.74	**59.26**
Slider	AM	63.68	**69.95**	56.54	**62.49**
	HM	62.77	**68.98**	56.27	**62.28**
Valve	AM	50.59	**54.53**	50.33	**50.72**
	HM	50.38	**54.41**	50.29	**50.70**
Average	AM	59.48	**63.55**	54.13	**56.69**
	HM	55.05	**60.13**	53.76	**56.20**

TABLE II
OVERALL RESULTS BY USING THE PROPOSED (LNS) AND BASELINE (LMS) FEATURES IN TERMS OF AUC (%) AND PAUC (%) IN THE EVALUATION DATASET.

Machines	Sections	AUC		pAUC	
		LMS	Proposed	LMS	Proposed
Toy car	AM	59.20	**70.43**	56.91	**63.32**
	HM	57.07	**66.74**	56.46	**62.49**
Toy train	AM	44.73	**46.13**	**50.26**	49.09
	HM	**44.44**	43.53	**50.25**	49.06
Bearing	AM	44.79	**51.86**	50.23	**51.09**
	HM	43.20	**51.43**	50.17	**51.09**
Fan	AM	48.81	**50.90**	51.07	**51.46**
	HM	47.91	**50.54**	51.02	**51.43**
Gearbox	AM	51.63	**54.45**	50.40	**52.32**
	HM	50.40	**53.09**	50.40	**52.19**
Slider	AM	**49.16**	48.35	**50.61**	50.23
	HM	**44.52**	44.45	**50.56**	50.18
Valve	AM	45.60	**45.31**	49.65	**49.68**
	HM	45.15	**45.27**	49.62	**49.67**
Average	AM	49.13	**52.49**	51.31	**52.45**
	HM	47.14	**49.78**	51.13	**52.00**

both development and evaluation datasets are listed in Tables I and II, respectively. All results are shown in percentages, and the improved results are highlighted in bold.

The proposed LNS generally improved the performance in both development and evaluation datasets for most of the machines. The LNS obtained the highest improvement in the bearing of the development dataset and in the toy car of the evaluation dataset, which are relative improvements of 17.02 and 16.94 % for HM, respectively. By using the proposed LNS, averaged HMs improved from 55.05 to 60.13% in the development dataset and from 47.14 to 49.78% in the evaluation dataset, achieving relative improvements of 9.22 and 5.60% respectively.

There are two exceptions. The proposed LNS provided better performance in the fan of the evaluation dataset, but the performance degraded in the development dataset. In contrast, the LNS performed better in the slider of the development dataset even when degradation occurred in the evaluation dataset. This could be because the frequency-band importance is calculated independently with Eq. (1), which makes it difficult to consider the combined effects of each frequency component for all machines with the F-ratio. A more suitable quantification method could further improve performance.

Figure 3 shows a comparison of results between our proposed method with 83 other methods for fan and toy car. Each dot corresponds to a different system proposed in the DCASE2022 Challenge Task 2. There was a significant improvement by replacing the LMS of the baseline (blue dot) with the LNS extracted with our quantification results (red dot). The other state-of-the-art methods appeared to have higher performances; however, they employ heavy deep NNs containing hundreds of millions of parameters and pre-trained models.

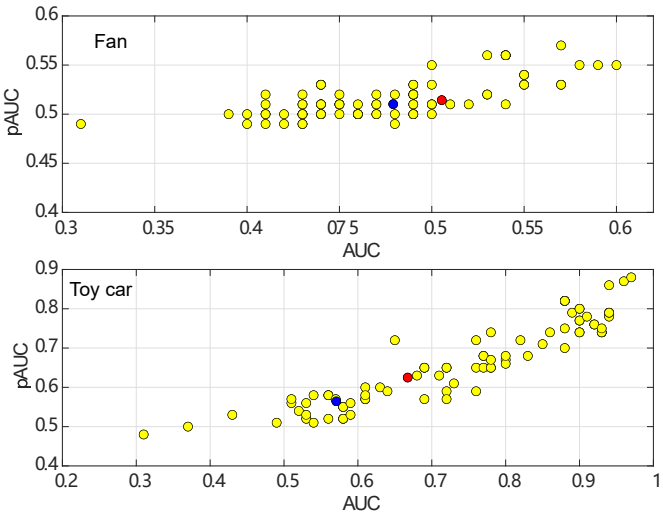

Fig. 3. Results in the evaluation dataset of DCASE2022 Challenge Task 2 using sounds recorded from the fan and toy car. Blue and red dots correspond to the baseline and proposed systems, respectively. The higher AUC and pAUC, the better performance.

VI. CONCLUSION

We quantified the importance of different frequencies for anomalous detection of seven types of machines using a data-driven statistical-based quantification method (machine-

wise F-ratio). We found that the discriminative features of each machine were encoded non-uniformly in the frequency domain. To highlight such important frequencies, we designed non-uniform FBs that have high resolutions in the frequencies with high F-ratios and used them to extract the LNS. The correctness of quantification results and effectiveness of the proposed LNS were verified in the DCASE2022 Challenge Task 2 with a simple AE-based detector. Compared with the LMS, the LNS achieved a relative improvement of 9.22 and 5.60% in development and evaluation datasets in terms of averaged HM of AUC, respectively. Future work will involve more sophisticated NN-based detectors to further improve the performance of an ASD system.

VII. Acknowledgements

This work was supported by JSPS-NSFC Bilateral Joint Research Projects/Seminars (JSJSBP120197416), Grant-in-Aid for Transformative Research Areas (A) (23H04344), SCOPE Program of Ministry of Internal Affairs and Communications (Grant Number: 201605002), and the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233).

References

[1] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?" in *Proc. 27th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE*, 2017, pp. 1–6.

[2] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," *arXiv preprint arXiv:2102.07820*, 2021

[3] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation," *arXiv preprint arXiv:2006.15321*, 2020.

[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," *arXiv preprint arXiv:2206.05876*, 2022.

[5] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. ICASSP. IEEE*, 2020, pp. 271–275.

[6] S. Kapka, "Id-conditioned auto-encoder for unsupervised anomaly detection," *arXiv preprint arXiv:2007.05314*, 2020.

[7] 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, 2018.

[8] T. Hayashi, T. Yoshimura, and Y. Adachi, "Conformer-based id-aware autoencoder for unsupervised anomalous sound detection," *DCASE2020 Challenge*, Tech. Rep., 2020.

[9] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a blstm network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58043–58055, 2018.

[10] T. Hayashi, T. Komatsu, R. Kondo, T. Toda, and K. Takeda, "Anomalous sound event detection based on wavenet," in *Proc. 26th EUSIPCO. IEEE*, 2018, pp. 2494–2498.

[11] E. Marchi, F. Vesperini, S. Squartini, B. Schuller et al., "Deep recurrent neural network-based autoencoders for acoustic novelty detection," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[12] K. Morita, T. Yano, and K. Tran, "Anomalous sound detection using cnn-based features by self supervised learning," *DCASE2021 Challenge*, Tech. Rep., 2021.

[13] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proc. 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 46–50.

[14] H. Van Truong, N. C. Hieu, P. N. Giao, and N. X. Phong, "Unsupervised detection of anomalous sound for machine condition monitoring using fully connected u-net." *Journal of ICT Research & Applications*, vol. 15, no. 1, 2021.

[15] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. IEEE ICASSP*, 2021, pp. 336–340.

[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[17] A. R. Webb, Statistical pattern recognition. John Wiley & Sons, 2003.

[18] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication*, vol. 50, no. 4, pp. 312–322, 2008.

[19] K. Li, X. Lu, M. Akagi, J. Dang, S. Li, and M. Unoki, "Relationship between speakers' physiological structure and acoustic speech signals: Data-driven study based on frequency-wise attentional neural network," in *Proc. EUSIPCO. IEEE*, 2022, pp. 379–383.

[20] Y. Zhou, Y. Sun, J. Li, J. Zhang, and Y. Yan, "Physiologically-inspired feature extraction for emotion recognition," in *Proc. Tenth Annual Conference of the International Speech Communication Association*. Citeseer, 2009.

[21] S. Hyon, J. Dang, H. Feng, H. Wang, and K. Honda, "Detection of speaker individual information using a phoneme effect suppression method," *Speech Communication*, vol. 57, pp. 87–100, 2014.

[22] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," In *arXiv e-prints: 2205.13879*, 2022.

[23] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proc. 6th DCASE*, Barcelona, Spain, November 2021, pp. 1–5.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.