Anomalous Machine Sound Detection Based on Time Domain Gammatone Spectrogram Feature and IDNN Model

Primanda Adyatma Hafiz[1], Candy Olivia Mawalim[2], Dessi Puji Lestari[1], Sakriani Sakti[3], and Masashi Unoki[2]

[1] Bandung Institute of Technology, Bandung, Indonesia
E-mail: {13520022@std, dessipuji@staff}.stei.itb.ac.id

[2] Japan Advanced Institute of Science and Technology, Nomi, Japan
E-mail: {candylim, unoki}@jaist.ac.jp

[3] Nara Institute of Science and Technology, Nara, Japan
E-mail: ssakti@is.naist.jp

***Abstract*—Anomalous sound detection (ASD) systems distinguish normal and abnormal machinery conditions on the basis of sound. While most ASD systems rely on the log Mel spectrogram, it lacks sufficient frequency resolution and performs suboptimally for rapidly changing sounds. Alternatively, the Gammatone spectrogram, extracted using a time domain Gammatone filterbank, offers enhanced spectrogram resolution. This study investigates the effectiveness of the time domain Gammatone spectrogram for ASD. To optimally learn the time domain Gammatone spectrogram features, an Interpolation Deep Neural Network (IDNN) model was proposed as the detection model. This model detects nonstationary sound frames highly reliably. An evaluation was conducted using MIMII dataset with area under receiver operating characteristic curve (ROC AUC) as the metric. Experimental results showed that our proposed method achieved ROC AUC of 92.5%, outperforming the log Mel spectrogram feature by 5.9 percentage points.**

***Index Terms*—Anomalous sound detection, time domain Gammatone filterbank, IDNN, spectrogram**

I. Introduction

In anomalous sound detection (ASD), unsupervised learning methods are often used to assess machine conditions. With this approach, a model can distinguish a machine's typical patterns using solely data of machine sound's in normal conditions for training [1]. Consequently, detecting anomalies in machines does not necessarily require data of machine sounds in anomalous or defective conditions, which is difficult to obtain [2].

Most studies on ASD utilize deterministic acoustic features, primarily spectrogram-based features [3] due to their rich representations of time, frequency, and intensity, which make them ideal for neural network-based models [4]. The log Mel spectrogram, which is extracted in the frequency domain, has been used in various studies on ASD and is mainly modeled using various neural network-based models [5]. However, the Log Mel spectrogram has limitations in classifying machine conditions for some machines [6]. Additionally, using short-term Fourier transform (STFT) to extract spectrogram in the frequency domain results in suboptimal resolution due to the inherent weaknesses of the STFT method [7][8].

Meanwhile, the Gammatone spectrogram, another variation of the spectrogram, closely mimics the human auditory system's stimulus response, with its parameters also derived from psychoacoustic experiments [9]. Therefore, the Gammatone spectrogram better simulates a human experts' ability to distinguish machine sounds. While most spectrograms are frequency domain-based [10], the Gammatone spectrogram can be directly calculated from the time domain representation. This is achieved by using a time domain Gammatone filterbank to transform the raw waveform directly [11].

Modeling spectrograms requires a model capable of processing 2-dimensional data. Employing a standard autoencoder (AE) model alone often yields suboptimal results due to challenges in accurately reconstructing nonstationary frames [6]. This issue can be addressed by utilizing an Interpolation Deep Neural Network (IDNN) model, which predicts only the central frame rather than all frames. By predicting only the central frame, IDNN has been proven to work better especially for modelling nonstationary sound [12].

Another study reported a UNet model can outperform the IDNN model in terms of ASD form some machines [13]. UNet architecture has the potential to be used in IDNN since it has some advantages over the IDNN model that can potentially result in a complementary effect.

This paper aims to propose a novel unsupervised ASD method employing a time domain Gammatone spectrogram. The IDNN model, enhanced by an overlapping frame technique, is applied to enhance model performance. Additionally, the IDNN model is also tested with both AE and UNet architectures to determine the most optimal design. The study findings offer insights into the effectiveness of various spectrograms, particularly when modeled using the IDNN approach for ASD.

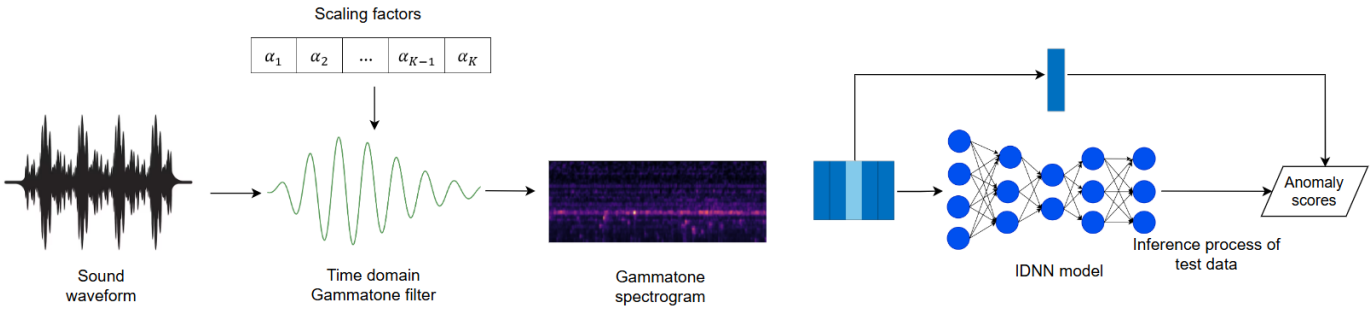

Fig. 1: Proposed method using time domain Gammatone spectrogram and IDNN model for anomalous sound detection

II. Anomalous Sound Detection

ASD involves determining whether the sound emitted from a machine is normal or anomalous. One challenge in this task is the scarcity of anomalous data for training, prompting the adoption of unsupervised learning approaches, which can be constructed using only normal training data [14].

Deterministic acoustic features are commonly favored in ASD systems due to their potential for better generalization, especially when data quantity is limited [15]. Spectrograms, which offer rich acoustic information from sound signals, are widely utilized in ASD systems [5]. One type of spectrogram is the Gammatone spectrogram derived from a Gammatone filterbank. Studies in automatic speech recognition have highlighted the effectiveness of Gammatone filterbanks over mel filterbanks in terms of noise robustness [16]. Although one method [17] employs Gammatone spectrograms extracted in the frequency domain, the use of STFT may potentially reduce resolution.

Qi et al. reported spectrograms derived directly from the time domain offer better representation than those derived through intermediate processes like STFT in the frequency domain. This is caused by frequencies being unnecessarily approximated during STFT calculation, which may result in the loss of some information [7]. Additionally, STFT also lacks resolution for representing sound with fast temporal changes, so it is not particularly suitable for nonstationary sound [8]. Therefore, with spectrograms extracted from the frequency domain, anomalous sounds are potentially harder to detect in valves than in other machines due to nonstationary and sparse characteristics in the sound signal [18].

Regarding the model, unsupervised ASD typically involves training solely on normal data and treating anomalous data as outliers. A basic implementation is the standard AE model, which learns to reconstruct only normal input data [19]. Consequently, the reconstruction of anomalous data may not be as accurate. Some studies have employed self-supervised learning techniques, learning machine label IDs and flagging data as anomalous if there is a significant deviation in predicted IDs [20]. However, self-supervised techniques often perform less unstable across different machines than unsupervised models [21] and rely on data augmentation methods.

III. Proposed Method

The proposed method comprises two primary processes: extracting the time domain Gammatone spectrogram and training the IDNN model for ASD, as illustrated in Fig. 1. During training, the model exclusively utilizes normal data, resulting in higher reconstruction error values for anomalous data, effectively representing the anomaly score.

A. Time Domain Gammatone Spectrogram Feature

A Gammatone spectrogram is a time-frequency domain feature that is calculated using the Gammatone filterbank. The Gammatone filterbank is a well-known set of filters used to simulate the response of the basilar membrane [9]. The time domain Gammatone filterbank uses a gamma function on multiple centers of frequency (f_c) to directly transform the raw waveform.

The impulse response of the Gammatone Filterbank at a center frequency f_c is defined as

$$g(t) = t^{n-1} e^{-2\pi b \mathrm{ERB}(\mathrm{f_c})\mathrm{t}} e^{j2\pi f_c t} \tag{1}$$

where $t \geq 0$ is the time in seconds, n is the filter order, and b is the bandwidth coefficient. The Gammatone spectrogram uses Equal Rectangular Bandwith Scale (ERB) to define the nonlinear spacing of the Gammatone frequency band [22]. ERB has finer frequency resolution than mel scale at lower frequency [23]. Therefore, the Gammatone spectrogram can prevent the loss of information at low frequency which in most cases is important for machine sounds. ERB at frequency center f_c is defined as

$$\mathrm{ERB}(\mathrm{f_c}) = 24.7 + 0.108\mathrm{f_c} \tag{2}$$

The Gammatone filterbank consists of K Gammatone filters $g^{(k)}(t)$ with different center frequencies. The output of the filterbank $X_k(t)$ from an input signal $x(t)$ can be defined as

$$X_k(t) = x(t) * g^{(k)}(t) \tag{3}$$

The Gammatone filterbank can be implemented using a wavelet transform where the mother wavelet is $\psi(t) = g(t)$ [24]. Then, with $\alpha > 1$, the k-th filter $g^{(k)}(t)$ can be defined by scaling $\psi(t)$ with a factor a_k of t, as

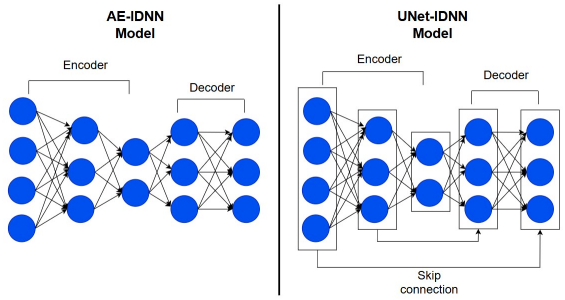

Fig. 2: Architectures of AE-IDNN and UNet-IDNN model

$$g^{(k)}(t) = \psi(\alpha_k t) \tag{4}$$

$$\alpha_k = \alpha^{\frac{2k}{K-1}-1} \tag{5}$$

The process of extracting and modeling the Gammatone spectrogram is shown in Fig 1.

B. Interpolation Deep Neural Network (IDNN) Model with Autoencoder and UNet Architecture

The IDNN model is a neural network-based model that is an improvement over the conventional approach of using a standard AE model to reconstruct the spectrogram input. IDNN overcomes the difficulty of detecting edge frames in standard AE, which is especially difficult for nonstationary sounds [12].

Instead of reconstructing all input, IDNN uses the center frame prediction method. This means that IDNN constructs the center frame on the basis of the input of left and right frames. Therefore, IDNN receives inputs of $[x_1, ..., x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}+1}, ..., x_n]$ frames and predicts the center frame ($x_{\frac{n+1}{2}}$). The loss function of IDNN is defined as

$$L(x_{\frac{n+1}{2}} | D(E([x_1, ..., x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}+1}, ..., x_n])) \tag{6}$$

where L defines the algorithm of the loss function used in IDNN.

IDNN performs better at predicting nonstationary frames using the center frame prediction method [12]. Thus, to increase the quantity of training data, instead of predicting only $\frac{N}{K}$ center frames with N as the total number of frames and K as the number of frames predicted each time, the proposed method uses $N - K + 1$ center frames with some frames overlapping in the data. Therefore, the i_{th} processed data are $[x_i, x_{i+1}, ..., x_{i+K-1}]$ frames instead of $[x_{(i-1)K)}, x_{(i-1)K)+1}, ..., x_{iK-1}]$ frames. The method is better able to handle the limited quantity of data and the rare occurence of anomalous conditions.

IDNN is implemented with AE architecture (AE-IDNN), which is its standard architecture [12] and UNet architecture (UNet-IDNN), which has the potential to complement the use of the center frame prediction method [13]. With AE architecture, IDNN consists of an encoder and decoder, where the decoder aims to predict the center frame. Meanwhile, with UNet architecture, IDNN consists of an encoder, a decoder, and additional skip connections. With the skip connection, each layer is connected with not only its successive layers because the skip connection creates direct links between layers of the encoder and corresponding layers of the decoder, thus providing more information to the model [25]. The architectures of the AE-IDNN and UNet-IDNN model are depicted in Fig. 2

TABLE I: Normal and anomalous data distribution in MIMII dataset

Machine		Normal data	Anomalous data
Slider	ID 00	1068	356
	ID 02	1068	267
	ID 04	534	178
	ID 06	534	89
Fan	ID 00	1011	407
	ID 02	1016	359
	ID 04	1033	348
	ID 06	1015	361
Valve	ID 00	991	119
	ID 02	708	120
	ID 04	1000	120
	ID 06	992	120
Pump	ID 00	1006	143
	ID 02	1005	111
	ID 04	702	100
	ID 06	1036	102
Total		14719	3400

IV. Experiment

A. Datasets

We used real machinery sounds data from the malfunctioning industrial machine investigation and inspection (MIMII) dataset for evaluation [18] at a 6 dB signal-to-noise ratio (SNR) because the recorder is assumed to be located closer to the target machine. This dataset consists of sounds from four different actual machines including sliders, fans, valves, and pumps augmented with real factory noise. Generally, each machine can be divided into two main categories by its emitted sounds' characteristics including stationary sound (fans and pumps) and nonstationary sound (valves and sliders) [12][18][26]. The MIMII dataset is already used in the DCASE challenge for ASD task [14]. It also contains actual machinery sounds in various anomalous conditions.

Each machine type consists of four machine IDs, with the distribution shown in Table I. For each machine ID, the model is trained separately using the unsupervised method. The inference process uses all anomalous data and normal data with the same amount as anomalous data, whereas the training process uses the remaining normal data.

B. Evaluation Metrics

The area under the receiver operating characteristic curve (ROC AUC) is used as the evaluation metric. It is calculated

TABLE II: Performance comparison of log Mel spectrogram and time domain Gammatone spectrogram feature across different models

Machine		Log Mel Spectrogram				Time Domain Gammatone Spectrogram			
		AE	UNet	AE-IDNN	UNet-IDNN	AE	UNet	AE-IDNN	UNet-IDNN
Slider	ID 00	0.991	0.989	0.958	0.992	1.000	0.997	0.997	0.997
	ID 02	0.893	0.947	0.916	0.984	0.690	0.715	0.822	0.839
	ID 04	0.744	0.799	0.786	0.875	0.906	0.940	0.984	0.954
	ID 06	0.656	0.653	0.668	0.739	0.976	0.997	1.000	0.999
	Average	0.821	0.847	0.832	0.898	0.893	0.912	0.951	0.947
Fan	ID 00	0.763	0.809	0.716	0.765	0.803	0.844	0.855	0.884
	ID 02	0.937	0.947	0.913	0.903	0.911	0.876	0.939	0.917
	ID 04	0.923	0.920	0.891	0.943	0.996	0.994	0.987	0.991
	ID 06	0.981	0.988	0.975	0.973	0.996	1.000	0.995	1.000
	Average	0.901	0.916	0.874	0.896	0.927	0.929	0.944	0.948
Valve	ID 00	0.550	0.656	0.702	0.847	0.450	0.463	0.922	0.947
	ID 02	0.622	0.682	0.704	0.887	0.929	0.957	1.000	1.000
	ID 04	0.602	0.708	0.665	0.932	0.698	0.845	0.946	0.916
	ID 06	0.666	0.696	0.732	0.900	0.523	0.639	0.854	0.836
	Average	0.610	0.686	0.701	0.892	0.650	0.729	0.931	0.925
Pump	ID 00	0.874	0.813	0.804	0.695	0.839	0.827	0.804	0.813
	ID 02	0.503	0.560	0.498	0.531	0.704	0.681	0.715	0.725
	ID 04	0.997	1.000	0.999	1.000	0.960	0.997	0.997	0.998
	ID 06	0.933	0.894	0.907	0.890	0.974	0.971	0.946	0.979
	Average	0.827	0.817	0.802	0.779	0.869	0.869	0.866	0.879
All	Average	0.790	0.816	0.802	0.866	0.835	0.859	0.923	0.925

by measuring the entire two-dimensional area underneath the entire ROC curve. The AUC is a metric used to determine the ability of a model to distinguish between classes [27] and is calculated using the following formula

$$\mathrm{AUC} = \frac{1}{\mathrm{N\_N_+}} \sum_{\mathrm{i}=1}^{\mathrm{N\_}} \sum_{\mathrm{j}=1}^{\mathrm{N_+}} \mathcal{H}(\mathcal{A}_\theta(\mathrm{x_j^+}) - \mathcal{A}_\theta(\mathrm{x_i^-})) \qquad (7)$$

N_- and N_+ subsequently represent the number of anomalous and normal data, x_i and x_j subsequently represent the current anomalous and normal data being processed, $\mathcal{H}(x)$ represents a function that return 1 if $x > 0$ and 0, otherwise, and $\mathcal{A}_\theta(x)$ represents the anomaly score of the data.

C. Experimental Configurations

The Gammatone spectrogram was extracted using the time domain Gammatone filterbank with $K = 64$ and $\alpha = 10$. For the mother wavelet $\psi(t)$, we use an 4^{th} order Gammatone filterbank ($n = 4$) with $b = 1.019$ and $f_c = 600$ Hz. The Gammatone spectrogram was also downsampled to reduce the temporal dimension to 310 frames for each 10-second sound data. Five frames were concatenated to produce a 320-dimensional input vector that was later fed into the model.

We compare the results of our system with a log Mel spectrogram feature that was calculated using librosa library implementation with $n_{\mathrm{FFT}} = 1024$, hop length $= 512$, and mel bands $= 64$. The spectrogram also used the concatenation of 5 frames to produce a 320-dimensional input vector that was later fed into the model.

For both spectrograms, IDNN received an input of $[x_1, x_2, x_4, x_5]$ frames and predicted the center frame (x_3) while minimizing the loss function of the mean squared error (MSE) loss. The hidden layers AE-IDNN consisted of $[64, 32, 16, 32, 64]$ neurons. Meanwhile, the hidden layers of UNet-IDNN consisted of $[64, 32, 16, 64, 128]$ neurons. UNet-IDNN also employed a batch normalization layer to complement the use of skip connections [28]. At the training stage, the time domain Gammatone spectrogram and log Mel spectrogram were trained at 200 epochs except for the log Mel spectrogram with AE-IDNN which was trained at 400 epochs because we observed that it could not provide an optimal solution at 200 epochs.

For comparison, we also used the standard AE model [18] and UNet model [13] to evaluate the generalizability of the proposed features. Both spectrograms were trained at 200 epochs for each model, and the reconstruction error was calculated using MSE.

V. Results

The results in Table II show that time domain Gammatone spectrogram performed better than the log Mel spectrogram

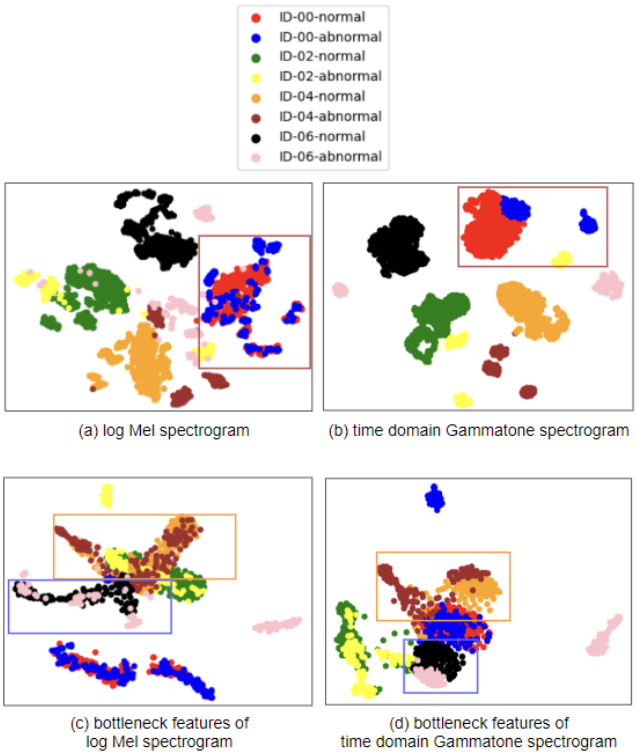

Fig. 3: The t-SNE visualization of the log Mel spectrogram, time domain Gammatone spectrogram, and bottleneck features of both spectrograms in UNet-IDNN model for the machine type fan. Different colors represent difference machine IDs and conditions. The normal and anomalous clusters are highlighted with colored contours for some machines that show a significant difference in the degree of separation between both features.

in all comparative models. Moreover, the time domain Gammatone spectrogram was also able to improve on the poor performance of the log Mel spectrogram on some machines. Hence, the time domain Gammatone spectrogram better distinguished normal and anomalous conditions than the log Mel spectrogram.

Even though using a conventional modeling approach with standard AE can already produce good results for the time domain Gammatone spectrogram, employing IDNN model further enhanced the results, especially on sliders and valves. The improvement of IDNN on sliders and valves was mostly caused by the characteristics of sliders and valves, which have more nonstationary characteristics than other machines such as fans and pumps. The most significant improvement lies in the valves with a more than 20-percentage point (pp) increase in AUC score compared to baseline model AE and UNet.

Additionally, UNet-IDNN performed slightly better than AE-IDNN with the time domain Gammatone spectrogram. Meanwhile, in the log Mel spectrogram, UNet-IDNN significantly improved over AE-IDNN. Thus, this shows the effectiveness of the skip connections component on IDNN for ASD.

Moreover, the use of the time domain Gammatone spectrogram yielded optimal results across both AE-IDNN and UNet-IDNN. Meanwhile, the log Mel spectrogram exhibited optimal result only in UNet-IDNN. This demonstrates that the advantages of the time domain Gammatone spectrogram led to more consistently optimal results across different IDNN architectures.

To further improve the performance of the UNet-IDNN model, we also employed hyperparameter tuning with a grid search method to find the most optimal structure and other parameters of the model [29]. However, hyperparameter tuning did not successfully improve the overall results.

VI. Discussion

When each feature was modelled using its best model which is UNet-IDNN, the time domain Gammatone spectrogram improved over log Mel spectrogram by 5.9 pp overall. The results also demonstrate the time domain Gammatone spectrogram's ability to model both stationary and nonstationary sound as the overall performance is better than log Mel spectrogram for every machie type in UNet-IDNN. The improvement in nonstationary machines, particularly valves, indicates the potential of the time domain Gammatone spectrogram for representing sounds with rapidly changing characteristics.

In addition to non-stationary sound, the time domain Gammatone spectrogram functions well in stationary machines such as fans. The effectiveness of this approach was evaluated via t-distributed stochastic neighbor embedding (t-SNE) cluster visualization of the latent features of the log Mel spectrogram and time domain Gammatone spectrogram as shown in Fig. 3(a) and 3(b). T-SNE distribution is chosen because of its ability to distinguish nonlinearly separable data which is similar to neural network model capability [30]. The figures show the ability of the time domain Gammatone spectrogram to distinguish normal (red) and anomalous (blue) conditions in fan ID 00 compared to the log Mel spectrogram.

Moreover, the time domain Gammatone spectrogram's performance is also complemented by the use of IDNN models. Fig. 3(c) and 3(d) show the t-SNE distribution of bottleneck features produced by the UNet-IDNN model for fans. In comparison to the log Mel spectrogram, bottleneck features of time domain Gammatone spectrogram can better separate normal and anomalous data. Hence, this indicates the compatibility between the time domain Gammatone spectrogram and neural network-based models.

However, the separation between normal and anomalous data is more distinguishable in the t-SNE distribution of the original features than that of the bottleneck features, as shown in Fig. 3. This indicates the inability of bottleneck features to solely represent spectrogram features, highlighting the importance of input from skip connections.

Although the time domain Gammatone spectrogram performs optimally, it has limitations especially in terms of computational complexity. Compared to STFT's [31], extracting the time domain Gammatone spectrogram requires

more computational resources due to the computation of an infinite impulse response (IIR) filter over multiple frequency bands. Therefore, this makes the time domain Gammatone spectrogram unsuited to real time ASD.

VII. Conclusion

This study proposed an unsupervised learning approach for anomalous sound detection using the time domain Gammatone spectrogram feature and modeled using Interpolation Deep Neural Network (IDNN) based models. Our proposed method demonstrated a significant improvement of 5.9 pp over the baseline feature, the log Mel spectrogram. Moreover, the time domain Gammatone spectrogram performed better overall than the log Mel spectrogram in all comparative models. Utilizing the IDNN model further enhanced the effectiveness of the time domain Gammatone spectrogram feature, particularly IDNN with UNet architecture. This study underscores the superiority of the time domain Gammatone spectrogram in representing machine sounds compared to the log Mel spectrogram feature, and highlights the effectiveness of the IDNN model for spectrogram-based features.

Acknowledgment

This work was supported by the JST Sakura Science Exchange Program, Grant-in-Aid for Transformative Research Areas (A) (23H04344), and JSPS KAKENHI grant (No. 22K21304).

References

[1] J. Zipfel, F. Verworner, M. Fischer, U. Wieland, M. Kraus, and P. Zschech, "Anomaly detection for industrial quality assurance: A comparative evaluation of unsupervised deep learning models," *Computers & Industrial Engineering*, vol. 177, p. 109045, 2023.

[2] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[3] 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.

[4] S. Abbasi, M. Famouri, M. J. Shafiee, and A. Wong, "Outliernets: Highly compact deep autoencoder network architectures for on-device acoustic anomaly detection," *Sensors*, vol. 21, no. 14, p. 4805, 2021.

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

[6] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[7] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2013, pp. 305–308.

[8] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992.

[9] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.

[10] 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. 58 043–58 055, 2018.

[11] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[12] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 271–275.

[13] T. V. Hoang, H. C. Nguyen, and G. N. Pham, "Unsupervised detection of anomalous sound for machine condition monitoring using different auto-encoder methods," Tech. Rep, Tech. Rep., 2020.

[14] 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.

[15] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via tasnet," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 36–40.

[16] L. Deng and Y. Gao, "Gammachirp filter banks applied in roust speaker recognition based on gmm-ubm classifier." *Int. Arab J. Inf. Technol.*, vol. 17, no. 2, pp. 170–177, 2020.

[17] 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.

[18] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," *arXiv preprint arXiv:1909.09347*, 2019.

[19] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pp. 353–374, 2023.

[20] W. Junjie, W. Jiajun, C. Shengbing, S. Yong, and L. Mengyuan, "Anomaly sound detection system based on multi-dimensional attention module," 2023.

[21] J. Guan, Y. Liu, Q. Kong, F. Xiao, Q. Zhu, J. Tian, and W. Wang, "Transformer-based autoencoder with id constraint for unsupervised anomalous sound detection," *EURASIP journal on audio, speech, and music processing*, vol. 2023, no. 1, p. 42, 2023.

[22] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[23] W. Lambamo, R. Srinivasagan, and W. Jifara, "Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition," *applied sciences*, vol. 13, no. 1, p. 569, 2022.

[24] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, no. 3-4, pp. 261–279, 1999.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[26] B. A. Tama, M. Vania, I. Kim, and S. Lim, "An efficientnet-based weighted ensemble model for industrial machine malfunction detection using acoustic signals," *IEEE Access*, vol. 10, pp. 34 625–34 636, 2022.

[27] C. X. Ling, J. Huang, H. Zhang *et al.*, "Auc: a statistically consistent and more discriminating measure than accuracy," in *Ijcai*, vol. 3, 2003, pp. 519–524.

[28] A. Labatie, "Characterizing well-behaved vs. pathological deep neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3611–3621.

[29] S. Mezzah and A. Tari, "Practical hyperparameters tuning of convolutional neural networks for eeg emotional features classification," *Intelligent Systems with Applications*, vol. 18, p. 200212, 2023.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[31] W. T. Cochran, J. W. Cooley, D. L. Favin, H. D. Helms, R. A. Kaenel, W. W. Lang, G. C. Maling, D. E. Nelson, C. M. Rader, and P. D. Welch, "What is the fast fourier transform?" *Proceedings of the IEEE*, vol. 55, no. 10, pp. 1664–1674, 1967.