Anomalous Sound Detection Based on Machine Activity Detection

Tomoya Nishida, Kota Dohi, Takashi Endo, Masaaki Yamamoto, Yohei Kawaguchi
Research and Development Group, Hitachi, Ltd.,
1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan

***Abstract*—We have developed an unsupervised anomalous sound detection method for machine condition monitoring that utilizes an auxiliary task — detecting when the target machine is active. First, we train a model that detects machine activity by using normal data with machine activity labels and then use the activity-detection error as the anomaly score for a given sound clip if we have access to the ground-truth activity labels in the inference phase. If these labels are not available, the anomaly score is calculated through outlier detection on the embedding vectors obtained by the activity-detection model. Solving this auxiliary task enables the model to learn the difference between the target machine sounds and similar background noise, which makes it possible to identify small deviations in the target sounds. Experimental results showed that the proposed method achieves a better performance than the conventional method particularly when the environmental noise contains sounds similar to the target machine sound. In addition, the proposed method improved the anomaly-detection performance of the conventional method complementarily by means of an ensemble.**

***Index Terms*—Machine health monitoring, anomalous sound detection, self-supervised learning, machine activity detection**

I. INTRODUCTION

Anomalous sound detection (ASD) is a task to identify whether a given sound is normal or anomalous. Since mechanical failure often causes machines to emit anomalous sounds, ASD has attracted attention for its application to machine condition monitoring [1], an essential technology for artificial intelligence-based factory automation. ASD is typically conducted in an unsupervised manner, meaning that only normal sounds are used for training. This is because anomalous sounds occur in rare situations and are highly diverse, making them almost impossible to collect. Most methods for unsupervised ASD (UASD) first learn a model of the collected normal sounds [2]–[7]. They then calculate the anomaly score of an observed sound on the basis of how well the sound fits the learned model. The sound is identified as anomalous if the score exceeds a preset threshold.

UASD for machine condition monitoring is often conducted in factories under noisy conditions, where the environmental noise tends to degrade the performance since the difference between normal and anomalous sounds is relatively small. This phenomenon is more pronounced when the environmental noise is similar to the target machine sounds. Recent methods for solving the noise problem [8]–[13] have utilized models that classify the sounds of the target machine and those of other similar machines, in contrast to the conventional UASD methods, which enables them to distinguish minor deviations between normal and anomalous sounds. An ensemble of these methods with other UASD methods has exhibited a good performance. However, it is extremely labor-intensive to find other machines similar to the target machine and then to record those sounds as training data in practical situations.

In this paper, we propose a UASD method that does not require sounds of other machines similar to the target machine. To solve the UASD task, the proposed method utilizes a model trained to solve an auxiliary task of detecting when the target machine is active. First, we train an activity-detection model that estimates when the target machine is active. Then, in the inference phase, we calculate the anomaly score by using the activity-detection error of the activity-detection model. Since the activity-detection model is trained to distinguish the sounds of the target machine from environmental noise, it can detect anomalous sounds especially when environmental noise is similar to the sounds of the target machine. Moreover, to enable inference without ground-truth machine activity labels, we propose applying an outlier detection method to the embeddings extracted from the activity-detection model.

II. RELATED WORK

A. UASD methods

Various UASD methods have targeted machine condition monitoring. Most of them learn a model of the normal sounds and then detect sounds that deviate from the learned model as anomalous. Several models have been used for learning the normal model, including autoencoders (AEs) [2], variational autoencoders [3], long short-term memories [5], transformers [6], normalizing flows [4], and Gaussian mixture models (GMMs) [7]. With these models, the anomaly score $\mathcal{A}(\boldsymbol{x};\boldsymbol{\theta})$ for a given input $\boldsymbol{x}$ is calculated in the inference phase, where $\boldsymbol{\theta}$ is the parameter of the model. If $\mathcal{A}(\boldsymbol{x};\boldsymbol{\theta})$ exceeds a predefined threshold, input $\boldsymbol{x}$ is detected as anomalous. Extensions to AEs and VAEs have also been proposed, such as changing the reconstruction task to an interpolation task [14], which improves the detection performance for non-stationary sounds, or cascading various types of dereverberation methods before the model [15]. All of these methods except the final one [15] learn the normal model without distinguishing the target machine sounds from environmental noise, and as a result, the existence of environmental noise degrades the ASD performance.

B. Self-supervised classification-based ASD methods

In Task 2 of Challenges on Detection and Classification of Acoustic Scenes and Events (DCASE Challenge) 2020 and

2021 [16], [17], many methods based on learning classification models that can be interpreted as a variation of outlier exposure [18] performed well. These methods train a model that identifies the machine ID for a given audio clip (sounds of several different individuals were provided for each machine type, with the individuals tagged as IDs), and the classification error is used as the anomaly score [8]–[13]. A normalizing-flow-based method using data from multiple machine IDs has also been proposed [19]. These methods tend to achieve better performances than unsupervised methods thanks to learning a good decision boundary between normal and anomalous samples by using the sounds from other machine IDs as proxy outliers. However, they need the sound of each class of the classification task to be similar to achieve these results [19], [20]. While this is possible in competitions where sounds for multiple machine IDs are provided for each machine type, in practical use it is quite costly to find appropriate machines and record their sounds.

III. PROBLEM STATEMENT

In this paper, we tackle the UASD task, i.e., anomalous sound detection under the condition that only normal sounds are available in the training phase. Unlike DCASE 2020 and 2021 Challenge Task 2, we consider a case where sounds for different individuals of the same machine type are unavailable. We also make the following assumptions.

1) The target machine repeatedly starts and stops during sound recording. We call the time when the machine is running *active* and the time when it is stopped *inactive*. During the active time, both the target machine sound and environmental noise are recorded, while during the inactive time, only the noise is recorded.
2) The training data contains information about when the target machine started and stopped running (called *activity labels*). If machine activity can be automatically recorded, the activity labels will be available in the training and inference phases. Even if they are not automatically recorded, the activity labels are available in the training phase because they can be annotated by hand.

IV. PROPOSED METHOD

A. Basic concept

The basic concept of the proposed UASD method is to use a model trained to solve an auxiliary task of detecting when the target machine is active. We call this model the "activity-detection model" and train it by using normal sounds of the target machine with ground-truth activity labels. If the activity-detection model fails to detect the active time frames in the inference phase, we regard the sound clip as anomalous. The activity-detection model is expected to learn a good decision boundary between the normal sounds of the target machine and other sounds, including anomalous sounds. The proposed UASD method based on activity detection works especially well when the environmental noise is similar to the target-machine sound. This case is likely to occur in factories because many machines are often similar to the target machine in operation.

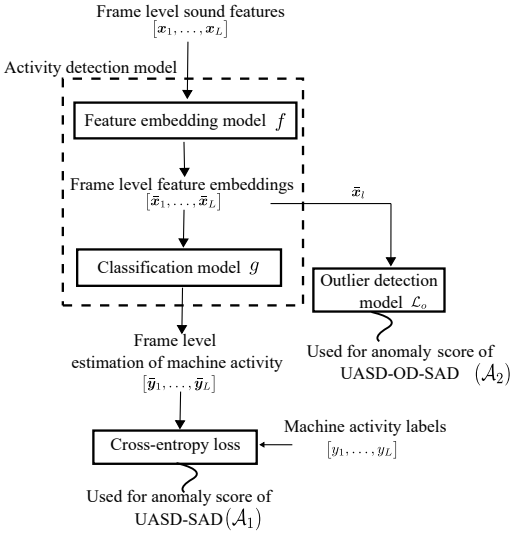


Fig. 1: Overview of proposed UASD-SAD and UASD-OD-SAD.

B. Case in which activity labels are available in inference

1) UASD based on supervised activity detection (UASD-SAD): As shown in Fig. 1, the activity-detection model consists of two components: a feature embedding model and a frame-wise classification model. First, F-dimensional frame-wise feature vectors (such as log-mel spectrograms) are extracted from sound clips and L consecutive features $\left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L\right]$, $\boldsymbol{x}_l \in \mathbb{R}^F$ $(l = 1, \ldots, L)$ are taken as input. The task is to estimate the activity labels $[y_1, \ldots, y_L]$, where $y_l = 1$ if the target machine was active in the l-th time frame and $y_l = 0$ if not. Note that only a few feature frames are input into this model, not the entire audio clip. The aim is to make activity detection difficult enough to be an auxiliary task for anomaly detection. If we input the whole audio clip, the activity will be detected precisely for both normal and anomalous sounds, and anomaly detection will not work.

The feature embedding model extracts L embedding vectors $\left[\bar{\boldsymbol{x}}_1, \cdots, \bar{\boldsymbol{x}}_L\right]$ from input as

$$\left[\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_L\right] = f\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L; \boldsymbol{\theta}_f\right), \tag{1}$$

where $\boldsymbol{\theta}_f$ denotes the parameters of the feature embedding model f. This model can be a convolutional neural network (CNN), a gated recurrent unit, or any other appropriate model. The classification model g then estimates the activity label of each frame from the feature embeddings, as

$$\bar{\boldsymbol{y}}_l = g(\bar{\boldsymbol{x}}_l; \boldsymbol{\theta}_g),\ l = 1, \ldots, L, \tag{2}$$

where $\bar{\boldsymbol{y}}_l \in (0, 1)^2$ and $\boldsymbol{\theta}_g$ denotes the parameters of the classification model g. The first and second component of $\bar{\boldsymbol{y}}_l$ can be interpreted as the posterior probability of the l-th time frame being inactive and active, respectively. Typically, g can

be given as a combination of a linear transform and a softmax function as

$$g(\bar{\boldsymbol{x}};\boldsymbol{\theta}_g) = \begin{bmatrix} \frac{\exp\left(\boldsymbol{w}_1^\mathsf{T}\bar{\boldsymbol{x}}\right)}{\exp\left(\boldsymbol{w}_1^\mathsf{T}\bar{\boldsymbol{x}}\right)+\exp\left(\boldsymbol{w}_2^\mathsf{T}\bar{\boldsymbol{x}}\right)} \\ \frac{\exp\left(\boldsymbol{w}_2^\mathsf{T}\bar{\boldsymbol{x}}\right)}{\exp\left(\boldsymbol{w}_1^\mathsf{T}\bar{\boldsymbol{x}}\right)+\exp\left(\boldsymbol{w}_2^\mathsf{T}\bar{\boldsymbol{x}}\right)} \end{bmatrix}, \tag{3}$$

where $\boldsymbol{\theta}_g = \{\boldsymbol{w}_1, \boldsymbol{w}_2\}$. Finally, the detection error is calculated as the cross entropy loss between $\left[\bar{\boldsymbol{y}}_1, \ldots, \bar{\boldsymbol{y}}_L\right]$ and the machine activity labels as

$$\mathcal{L}\left(\left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L\right]; \boldsymbol{\theta}\right) = -\sum_{l=1}^{L} \log\left(\left[\bar{\boldsymbol{y}}_l\right]_{y_l+1}\right), \tag{4}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_f, \boldsymbol{\theta}_g\}$ and $[\boldsymbol{y}]_i$ denotes the i-th component of $\boldsymbol{y}$. The detection error is used both as the cost function used for training the model and as the anomaly score for an obtained sound. Since this method utilizes a supervised learning task of activity detection, we refer to it as *UASD based on supervised activity detection (UASD-SAD)*.

2) Overall cost function and anomaly score for a sound clip: We describe the overall cost function for a given training dataset and the anomaly score for a given sound clip. Both are calculated using a sliding window to extract L consecutive frames of feature vectors, which are then used as the input of the activity-detection model.

Assume that the training data consists of K sound clips of normal data $\mathcal{D} = \{\boldsymbol{X}^{(k)}\}_{k=1}^{K}$, where each sound clips consists of $T_k (\geq L)$ frames of feature vectors: $\boldsymbol{X}^{(k)} = \left[\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_{T_k}^{(k)}\right]$. For the input of the activity-detection model, L consecutive feature vectors starting from index t are denoted as

$$\boldsymbol{X}_t^{(k)} = \left[\boldsymbol{x}_t^{(k)}, \ldots, \boldsymbol{x}_{t+L-1}^{(k)}\right]. \tag{5}$$

Using this notion, we define the overall cost function for training the activity-detection model as

$$\mathcal{L}_{\text{cost}}\left(\mathcal{D}; \boldsymbol{\theta}\right) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{T_k - L + 1} \sum_{t=1}^{T_k - L + 1} \mathcal{L}\left(\boldsymbol{X}_t^{(k)}; \boldsymbol{\theta}\right). \tag{6}$$

In the same way, we define the anomaly score for a given sound clip $\boldsymbol{X} = \left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\right]$ as

$$\mathcal{A}_1\left(\boldsymbol{X}; \boldsymbol{\theta}\right) = \frac{1}{T - L + 1} \sum_{t=1}^{T - L + 1} \mathcal{L}\left(\boldsymbol{X}_t; \boldsymbol{\theta}\right). \tag{7}$$

C. Case in which activity labels are unavailable in inference

When the machine activity is not automatically recorded, the anomaly score defined in (7) cannot be computed. To deal with this situation, we propose training an outlier detector with the embedding vectors (which are the outputs of f) of the training data after the activity-detection model has been trained. For example, we can use a GMM or an AE as an outlier detector. Here, the anomaly score for $\boldsymbol{X}$ is calculated as

$$\mathcal{A}_2\left(\boldsymbol{X}; \boldsymbol{\theta}_f, \boldsymbol{\theta}_o\right) = \frac{1}{(T - L + 1)L} \sum_{t=1}^{T - L + 1} \sum_{l=1}^{L} \mathcal{L}_o(\bar{\boldsymbol{x}}_l^t; \boldsymbol{\theta}_o), \tag{8}$$

TABLE I: Requirements for activity labels.

Method	Training	Inference
(i) UASD w/ activity labels	Yes	Yes
(ii) UASD-SAD	Yes	Yes
(iii) UASD w/o activity labels	No	No
(iv) UASD-OD-SAD	Yes	No

where $\bar{\boldsymbol{x}}_l^t$ denotes the l-th embedding vector of $f(\boldsymbol{X}_t, \boldsymbol{\theta}_f)$, $\boldsymbol{\theta}_o$ denotes the parameters of the outlier detector, and $\mathcal{L}_o(\bar{\boldsymbol{x}}; \boldsymbol{\theta}_o)$ denotes the anomaly score for the outlier detector given an embedding vector $\bar{\boldsymbol{x}}$. Here, $\bar{\boldsymbol{x}}_l^t$ is expected to be close to either $\boldsymbol{w}_1$ or $\boldsymbol{w}_2$ in (3). Then, if a feature vector at a time frame that includes anomalous sounds is provided to the activity-detection model, $\bar{\boldsymbol{x}}_l^t$ would be dissimilar to both vectors and the anomaly score for this embedding vector would be high. Thus, anomalous sounds will have high anomaly scores. In this way, UASD can be conducted without using activity labels in the inference phase. We refer to this method as *UASD based on outlier detection using supervised activity detection (UASD-OD-SAD)*.

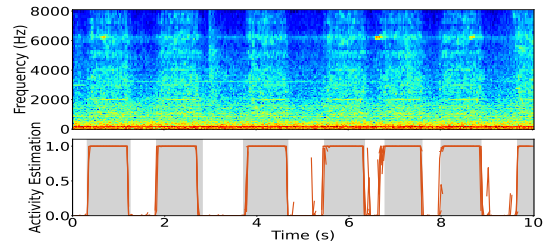
V. EXPERIMENTS

A. Experimental conditions

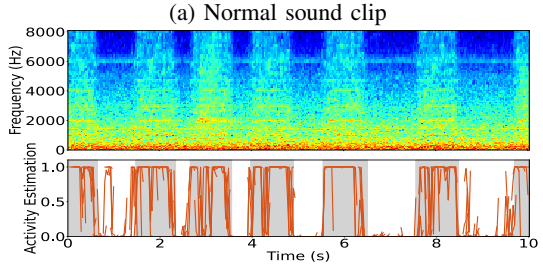
To investigate the effectiveness of the proposed method under noisy conditions, we compared its performance with that of a conventional UASD method using a machine sound dataset containing environmental noise.

For evaluation, we used the slide rail dataset included in the MIMII DUE dataset [21], as it satisfies our assumption that the input sounds contain both active and inactive sections of the target machine. Furthermore, slide rails are widely utilized in factories, and detecting their breakdown is critically important. We used the data in sections "00" and "01" in the development dataset, which contain different sounds. We annotated the active sections of the slide rail dataset for both the training and test data. To evaluate each method under low-signal-to-noise ratio (SNR) conditions, we mixed the original dataset with two types of environmental noise each recorded in different factories (Factory A and B). The SNR was between 6.0 dB and -12.0 dB, where the mixing procedure was the same as that previously reported [21]. Note that the original slide rail dataset already includes environmental noise, so the SNR given here is not the SNR between the clean slide rail sound and other noise. Instead, it is the SNR between the sounds of the original dataset and the additionally mixed in factory noise. Since the SNR of the original dataset was -12 dB, the actual SNR of the data used in our experiment was always below -12 dB. The input for each method was 128-dimensional log-mel spectrograms computed with a short-time Fourier transform frame size of 64 ms and a hop size of 50%.

We consider two scenarios: one in which activity labels are available in inference, and one in which they are not. For the first scenario, we compared the proposed method, UASD-SAD, with the conventional AE-based method using the activity labels (UASD w/ labels). In UASD w/ labels, by using the activity labels, the AE was trained only with the feature vectors of active time. In the evaluation phase, the mean reconstruction error of the feature vectors of active time was used as the



(a) Normal sound clip



(b) Anomalous sound clip

Fig. 2: Example of log-mel spectrogram (upper parts of (a), (b)) and activity detection (lower parts of (a), (b)) for UASD-SAD (section 00, Factory A noise, SNR $= 6.0$ dB). Gray areas show ground-truth activities, and each line represents the estimated activities for $L = 5$ consecutive frames.

anomaly score. For the second scenario, we compared the proposed method, UASD-OD-SAD, with the conventional AE-based method without the activity labels (UASD w/o labels). In UASD w/o labels, the AE was trained with all the feature vectors, including features of both active and inactive time. The anomaly score was also obtained by computing the mean reconstruction error of all the feature vectors. The architecture of the AE was the same as reported [21]: five fully connected layers for the encoder and the decoder with batch-norm layers located between every pair of layers. Each input contained five consecutive frames of extracted feature vectors concatenated to a single 640-dimensional vector.

We used a CNN-based architecture for the activity-detection model. The feature vectors of five consecutive frames were extracted to form a two-dimensional time-frequency representation. The feature vectors were input to a CNN layer and then by three residual CNN blocks. Each residual block consisted of two CNN layers and a residual connection, where each layer had 32 channels. The size and stride of the convolution kernel were 3×3 and 1. A fully connected layer then followed the CNN layers, which was applied to each feature of different time frames separately and identically. The architecture of this feature embedding model was designed so that the number of layers was approximately the same as the AE model, thus ensuring that the representation capacity of the two models would be close to each other. We used (3) for the activity-detection model and a GMM for the outlier-detection model. The number of mixture components in the GMM was 5. All neural networks were optimized by Adam [22]. The AEs were trained for 100 epochs, and the activity-detection models were trained for 20 epochs. Note that each method has different requirements for the activity labels, as shown in Table I.

For the first scenario, we evaluated an ensemble of UASD w/ labels and UASD-SAD. For the second scenario, we performed an ensemble of UASD w/o labels and UASD-OD-SAD. To ensemble different methods, the anomalous score of each model was first standardized and then summed up to calculate the overall anomaly score. The standardization of the anomaly score $\mathcal{A}(\boldsymbol{X})$ of a test data $\boldsymbol{X}$ was conducted by

$$\tilde{\mathcal{A}}(\boldsymbol{X}) = (\mathcal{A}(\boldsymbol{X}) - \mu) / \sqrt{\sigma^2 + \epsilon}, \tag{9}$$

where μ and σ^2 are the mean and the variance of $\mathcal{A}(\boldsymbol{X})$ of the training data, respectively, and ϵ is a constant positive value. We used $\epsilon = 1000$ for UASD-SAD since the variance of the anomaly score for the validation data tended to be much larger than σ^2 when we conducted cross-validation. We used $\epsilon = 0$ for all the other methods.

TABLE II: AUC values for anomaly detection (%).

(a) Methods that require activity labels in inference.

Method	Noise type / SNR [dB]	Factory A 6.0	Factory A 0	Factory A -6.0	Factory A -12.0	Factory B 6.0	Factory B 0	Factory B -6.0	Factory B -12.0
Section 00									
(i) UASD w/ labels		76.6	70.8	59.5	50.7	73.4	66.3	58.6	51.0
(ii) UASD-SAD		73.0	70.7	**62.8**	**58.7**	68.9	68.9	**68.8**	**55.1**
Ensemble of (i) and (ii)		**81.5**	**73.6**	58.3	51.5	**77.7**	**72.1**	62.7	52.1
Section 01									
(i) UASD w/ labels		**82.2**	77.7	72.4	**63.3**	**82.3**	76.7	67.8	**60.5**
(ii) UASD-SAD		81.6	73.9	63.5	53.7	80.0	74.3	64.2	53.2
Ensemble of (i) and (ii)		81.0	**81.5**	**74.9**	58.5	82.1	**82.6**	**72.3**	57.5
Average									
(i) UASD w/ labels		79.4	74.3	66.0	**57.0**	77.9	71.5	63.2	**55.8**
(ii) UASD-SAD		77.3	72.3	63.2	56.2	74.5	71.6	66.5	54.2
Ensemble of (i) and (ii)		**81.3**	**77.6**	**66.6**	55.0	**79.9**	**77.4**	**67.5**	54.8

(b) Methods that do not require activity labels in inference.

Method	Noise type / SNR [dB]	Factory A 6.0	Factory A 0	Factory A -6.0	Factory A -12.0	Factory B 6.0	Factory B 0	Factory B -6.0	Factory B -12.0
Section 00									
(iii) UASD w/o labels		70.3	64.9	55.8	51.7	70.2	61.5	55.4	51.2
(iv) UASD-OD-SAD		76.2	66.5	56.3	52.2	70.1	**68.7**	**64.7**	**51.7**
Ensemble of (iii) and (iv)		**77.5**	**68.3**	**57.0**	**52.3**	**74.6**	67.7	60.7	**51.7**
Section 01									
(iii) UASD w/o labels		80.2	74.0	65.9	**57.4**	79.5	72.0	63.8	**57.0**
(iv) UASD-OD-SAD		71.4	72.6	57.4	47.1	72.1	69.1	62.1	47.5
Ensemble of (iii) and (iv)		**80.5**	**78.3**	**68.2**	53.9	**81.0**	**78.1**	**68.2**	53.7
Average									
(iii) UASD w/o labels		75.3	69.5	60.9	**54.6**	74.9	66.8	59.6	**53.8**
(iv) UASD-OD-SAD		73.8	69.6	56.9	49.7	71.1	68.9	63.4	49.6
Ensemble of (iii) and (iv)		**79.0**	**73.3**	**62.6**	53.1	**77.8**	**72.9**	**64.5**	52.7

B. Results

Examples of the activity detection of UASD-SAD are provided in Fig. 2. As we can see, activity was detected correctly for normal sounds, while there were many errors in activity detection for anomalous sounds. Therefore, the proposed method is expected to detect anomalies based on the error of activity detection.

Table II shows the area under the receiver operating characteristic curve (AUC) for anomaly-detection performance. First, in most conditions, the methods that require the activity labels in the inference phase had higher AUC values than those that do not require the activity labels. Next, in many of the noise conditions in Section 00, UASD-SAD showed higher AUC values than UASD w/ label, but in Section 01, UASD w/ label showed higher AUC values. Also, in most of the noise conditions in Section 00, UASD-OD-SAD showed higher AUC values than UASD w/o label, but in Section 01, UASD w/o label showed higher AUC values. One of the reasons for the conflicting results obtained in Section 00

and Section 01 is most likely the difference of the similarity between the target machine sound and the environmental noise. The proposed method is expected to show high performance when the target machine sound is somewhat similar to the noise. On the other hand, when the target machine sound is not similar to the noise at all, anomaly detection cannot be performed well because the auxiliary task, activity detection, is too easy. In fact, the noise of factories A and B contained sounds similar to the slide rail in Section 00 but not to the slide rail in Section 01. Overall, the proposed methods achieved a better performance than the conventional methods when the noise was similar to the target machine sound, which is a condition that degraded the performance of the conventional methods substantially. This advantage is crucial in practical situations, since factories often run several similar machines in the same area. In this case, environmental noise tends to be similar to the target machine sound.

Also, the results showed that the ensembles achieved higher AUC values than the conventional methods for both sections 00 and 01, except for the SNR of -12.0 dB. These results indicate that the proposed method is also useful for improving the anomaly-detection performance of the conventional methods complementarily by means of an ensemble. It is also suggested that the proposed method does not contribute to performance improvement when the SNR is extremely low.

VI. CONCLUSION

We proposed a method for anomalous sound detection based on machine activity detection. The proposed method calculates the anomaly score based on the error of activity detection if the ground-truth activity labels are available in the inference phase. If these labels are not available, it performs outlier detection for the embeddings obtained in the activity-detection model. Experimental results indicate that the proposed method achieves a better performance than the conventional method particularly when the environmental noise contains sounds similar to the target machine sound, which is a crucial advantage in practical applications. In addition, the proposed method improved the anomaly-detection performance of the conventional method complementarily by means of an ensemble.

REFERENCES

[1] 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 Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 212–224, 2019.

[2] Y. Koizumi, S. Saito, M. Yamaguchi, S. Murata, and N. Harada, "Batch uniformization for minimizing maximum anomaly score of DNN-based anomaly detection in sounds," in *Proc. IEEE Int. Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2019, pp. 6–10.

[3] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 2366–2370.

[4] M. Yamaguchi, Y. Koizumi, and N. Harada, "AdaFlow: Domain-adaptive density estimator with application to anomaly detection and unpaired cross-domain translation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 3647–3651.

[5] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 1996–2000.

[6] T. Hayashi, T. Yoshimura, and Y. Adachi, "Conformer-based ID-aware autoencoder for unsupervised anomalous sound detection," Tech. Rep., Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE Challenge), 2020.

[7] H. Purohit, R. Tanabe, T. Endo, K. Suefusa, Y. Nikaido, and Y. Kawaguchi, "Deep autoencoding GMM-based unsupervised anomaly detection in acoustic signals and its hyper-parameter optimization," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 175–179.

[8] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, "Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation," Tech. Rep., Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE Challenge), 2020.

[9] P. Primus, "Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers," Tech. Rep., Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE Challenge), 2020.

[10] T. Inoue, P. Vinayavekhin, S. Morikuni, S. Wang, T. H. Trong, D. Wood, M. Tatsubori, and R. Tachibana, "Detection of anomalous sounds for machine condition monitoring using classification confidence," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 66–70.

[11] J. Lopez, G. Stemmer, and P. Lopez-Meyer, "Ensemble of complementary anomaly detectors under domain shifted conditions," Tech. Rep., Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE Challenge), 2021.

[12] K. Morita, T. Yano, and K. Tran, "Anomalous sound detection using CNN-based features by self supervised learning," Tech. Rep., Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE Challenge), 2021.

[13] K. Wilkinghoff, "Sub-Cluster AdaCos: Learning representations for anomalous sound detection," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2021.

[14] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 271–275.

[15] Y. Kawaguchi, R. Tanabe, T. Endo, K. Ichige, and K. Hamada, "Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 865–869.

[16] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 Challenge Task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 81–85.

[17] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 Challenge Task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv:2106.04492*, 2021.

[18] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019.

[19] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 336–340.

[20] P. Primus, V. Haunschmid, P. Praher, and G. Widmer, "Anomalous sound detection as a simple binary classification problem with careful selection of proxy outlier examples," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 170–174.

[21] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," *arXiv preprint arXiv:2105.02702*, 2021.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.