Timbre-Based Anomaly Explanation without Anomalous Training Data

Tomoya Nishida, Harsh Purohit, Kota Dohi, Takashi Endo, Yohei Kawaguchi
Research and Development Group, Hitachi, Ltd.

***Abstract*—This paper presents a novel framework for explaining anomalous machine sounds in the context of unsupervised anomalous sound detection (UASD). While UASD techniques have been widely studied, a key challenge remains: understanding how anomalous sounds differ from normal sounds to better support machine condition monitoring. Existing methods for describing sound differences rely on anomalous sounds for training, which is impractical in real-world scenarios where such data is typically unavailable. To address this limitation, we propose a new framework that explains anomalous sound differences by a pre-selected list of timbre-related words, such as brightness and boominess, instead of free-form text captions. The relationship between such words and a given sound has been objectively modeled as "timbral metrics" through psychoacoustical research, allowing the estimation of changes in timbre without the need for training machine learning models on anomalous sounds. Furthermore, to handle variations in normal training data, we propose a method that uses a k-nearest neighbors approach in the audio embedding space to measure how an anomalous sample's timbre differs from normal samples, while simultaneously performing UASD. We evaluated the proposed method on the MIMII DG dataset, demonstrating its effectiveness in explaining anomalous sounds while conducting UASD at the same time.**

***Index Terms*—Anomalous sound detection, Timbral attributes**

I. Introduction

Anomalous sound detection (ASD) [1] is a task that detects sounds that are "not normal". Applying this to sounds emitted from machines can lead to detecting mechanical failures, which helps monitoring machine condition. Since anomalous sounds are challenging to collect, ASD is often challenged in scenarios where only normal sounds are available for training, known as Unsupervised ASD (UASD) [1]–[8].

(U)ASD only detects the condition (normal or anomalous) of a given sound without specifying how the anomalous sound differs from normal ones. Consequently, further analysis is necessary to determine the cause of the anomaly and whether repairs are needed. Identifying how the anomalous sound differs from normal sounds can make this analysis easier, as it may indicate the type of machine malfunction. We refer to such differences observed in anomalous sounds as the "anomalous difference". In relation to such motivation, [9] proposes a task that explains the differences between two sounds by training a model to generate captions describing how normal sounds differ from anomalous ones. A similar task was also explored in [10] for general sound events. However, these methods cannot be used in the typical UASD problem setting for the following reasons: 1) They require normal and anomalous sound pairs and ground truth difference captions for training. This is impractical because anomalous samples are unavailable in UASD and creating captions for paired data is time-consuming. 2) These methods only describe differences between two audio samples. UASD focuses on detecting whether a sample deviates from the entire normal data distribution, which means that anomalous differences should also be explained based on how the anomalous sample differs from the whole normal data distribution. 3) They focus only on generating captions for anomalous differences without performing UASD, whereas it is preferable for the explanations to align with the UASD results.

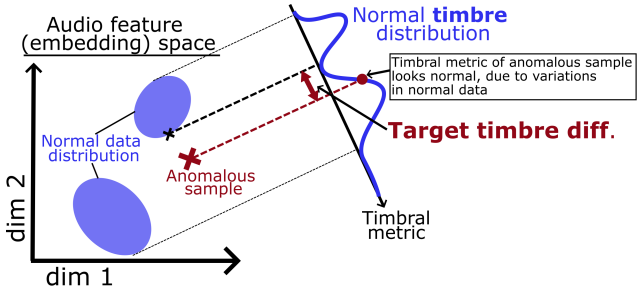

Fig. 1: Illustration of aim of proposed method. When comparing timbral metric values of anomalous sample to whole normal data, differences cannot be determined (red dot in normal timbre distribution). By comparing timbral metric only with neighbor normal samples in feature space, timbre differences can be determined (Target timbre diff.).

To address these issues, we propose a new strategy for anomalous difference explanation that suits the UASD problem setting and an anomalous difference explanation method that can be conducted along with UASD. **Specifically, to solve issue (1), we introduce an anomalous difference explanation framework that explains differences solely in pre-selected terms related to timbre, such as sharpness or boominess, instead of using free-form text captions (Contribution 1)**. The goal of this framework is to determine whether the sound impression represented by each timbre-related term has been reinforced or weakened. For instance, whether the sharpness has increased or the boominess has decreased, and so on. Objective metrics that quantify the degree of such impressions in sounds have been developed through psycho-acoustical research in the literature [11], [12], which enables us to infer the change in these impressions without model training. We term such changes in the impression of a sound "timbre difference", and term this framework "timbre difference capturing".

Furthermore, to ensure robustness against variations in

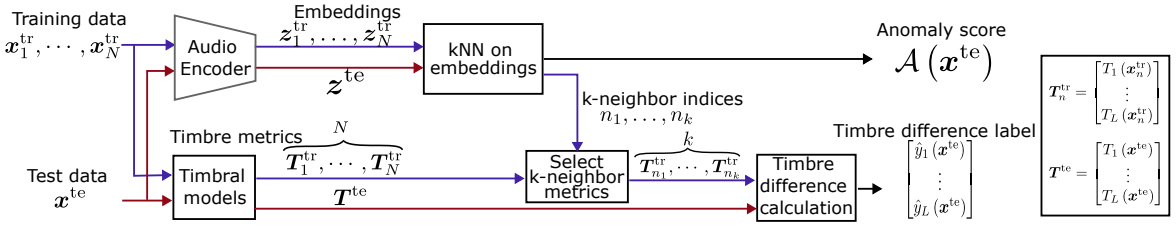

Fig. 2: Overview of proposed joint UASD and timbre difference capturing method in inference phase.

normal training data, **we further propose a method that performs timbre difference capturing by comparing the test sound with normal sounds most similar to it found in an embedding space (Contribution 2).** Finally, we created a dataset with ground truth timbre differences and evaluated the effectiveness of the proposed method.

II. Proposed Task: Timbre Difference Capturing

In this section, we propose a new task for explaining anomalous differences that suites the UASD problem setting. As in UASD settings [13]–[15], we only have normal data $\{\boldsymbol{x}_n^{(\mathrm{tr})}\}_{n=1}^{n=N}$ for training, possibly with optional metadata (e.g., machine speed). During inference, the system sees both normal and anomalous sounds without any extra information, and first must decide if a sound is normal or anomalous.

If a sample is decided as anomalous, the system has to explain how it is anomalous. Since anomalous data are not available for training, existing sound difference captioning methods [9], [10] cannot be applied. To address this, we introduce a novel framework "timbre difference capturing", which describes anomalous differences by pre-selected timbre-related terms rather than using free-form captions. For the timbre-related terms, we especially use words that are known as "timbral attributes" [16], which are adjectival words such as sharpness or boominess that decompose the various aspects of timbre. Research has developed mathematically formulated models that quantify these timbral attributes as objective metrics called "timbral metrics" [11], [12], which enables us to identify how strong humans will perceive that attribute without supervised machine learning. Additionally, timbre-related features have been applied for detecting machine malfunctions [17]–[19], demonstrating their relevance in detecting machine malfunctions. Still, these methods have not tried to directly explain anomalous differences using timbre, which will be the scope of this paper.

We predefine $L = 5$ timbral attributes relevant to machine failures, following [19]: **1) Sharpness:** Sharp or shrill sensation, **2) Roughness:** Buzzing, raspy sound quality, **3) Boominess:** Booming sensation, often perceived as low-pitch vibration, **4) Brightness:** Bright sensation, **5) Depth:** Emphasized low-frequency component. These attributes are modeled as timbral metrics in [12], which we use here. Using these attributes, the goal of timbre difference capturing is defined as follows: **Given an anomalous test sample $\boldsymbol{x}^{(\mathrm{te})}$, determine whether each attribute l ($l = 1, \ldots, L$) has increased, decreased, or remained unchanged due to the anomaly.** For example, if a machine malfunction causes an additional buzzing sound, the Roughness attribute may increase while other attributes remain unchanged. Here, "due to the anomaly" specifically refers to changes in timbre caused by machine malfunctions. Any timbre variations resulting from other factors, such as background noise or the machine's operational status, should be excluded. We represent the timbre differences by labels $y_l \in \{1, 0, -1\}$ $(1 \leq l \leq L)$, where $1, 0, -1$ stands for increased, no change, and decreased, respectively.

While timbral metrics for both normal and anomalous data can be computed directly, y_l cannot be estimated directly due to interference from factors like background noise or normal sound variations. The proposed method addresses this issue without requiring anomalous samples during training. Finally, although using a limited set of timbral attributes is less flexible than free-form captions, it still offers valuable insights into how anomalous sounds differ from normal ones. Extending beyond these attributes is left for future work.

III. Proposed Method

A. Overview

Although we can directly compute timbral metrics for both normal training data and anomalous inference data [12], this alone does not identify the true timbre difference. This is because machine sounds vary due to factors such as machine's operational modes, recording conditions, or noise, and timbre also varies due to such factors. To address this, we propose a method that focuses on extracting how the anomalous sound differs from normal sounds occured under the same conditions, ruling out irrelevant variations. Specifically, we assume that normal samples recorded under identical conditions as the anomalous sample are the samples that are most similar to that anomalous sample. By comparing the anomalous sample only with those similar normal samples, we can estimate its timbre difference more accurately. To find these similar samples, we use an audio encoder to extract embeddings for each sample, then identify the nearest neighbor normal samples in the embedding space. We illustrate this concept in Fig. 1.

Fig. 2 summarizes the proposed method. The method first extracts embeddings of given audio samples by an arbitrary audio encoder and conducts timbre difference capturing along with UASD based on those embeddings. Here, the embeddings are used to detect the k-nearest neighbor (knn) normal samples of the test data from $\{\boldsymbol{x}_n^{(\mathrm{tr})}\}_{n=1}^{N}$. The audio encoder can be a pre-trained model, trained from scratch, or a pre-trained model with fine-tuning. Since the choice of audio encoders are arbitrary, we omit explanations for this part. In the following

TABLE I: Statistics of ground truth timbre difference labels for $t' = 0.05$. #g, #u, r denote number of conditions, unique timbre difference label vectors, and labels of each values.

Section	section 00			section 01			section 02		
Machine	# g	# u	r $(-1/0/1)$	# g	# u	r $(-1/0/1)$	# g	# u	r $(-1/0/1)$
Bearing	26	23	32/49/34	32	19	16/44/35	4	1	2/3/0
Fan	3	2	1/6/3	7	6	7/6/17	3	3	5/9/1
Gearbox	26	17	14/28/43	23	13	6/15/44	6	6	8/9/13
Slider	26	11	24/14/17	26	10	26/15/9	6	4	4/11/5
Valve	4	2	2/7/1	8	8	3/29/8	7	4	9/2/9

subsections, we explain how UASD and timbre difference capturing are conducted in inference phase.

B. UASD

We denote $E(\boldsymbol{x})$ as the embedding obtained by the audio encoder $E(\cdot)$. Then, the anomaly score of a test sample $\boldsymbol{x}^{\text{te}}$ is given as the knn distance between $\boldsymbol{x}^{\text{te}}$ and $\{\boldsymbol{x}_n^{(\text{tr})}\}_{n=1}^{n=N}$ measured in the embedding space, which is

$$\mathcal{A}\left(\boldsymbol{x}^{\text{te}}\right) = \frac{1}{k}\sum_{i=1}^{k} d\left(E\left(\boldsymbol{x}^{\text{te}}\right), E\left(\boldsymbol{x}_{n_i}^{\text{tr}}\right)\right), \tag{1}$$

where $d(\cdot,\cdot)$ measures the distances between the given embeddings, such as the Cosine distance. n_i $(i = 1,\ldots,k)$ is the index of the i-th nearest neighbor training sample $\boldsymbol{x}_{n_i}^{\text{tr}}$, based on $d(\cdot,\cdot)$. This strategy of combining an audio encoder with a knn-based anomaly score calculator has been used in various UASD methods [13], [14], including methods used in top rankings of the latest UASD competitions [8], [20]. This means that the proposed method can be used to extend such methods for timbre difference capturing, which is a further advantage of the proposed method.

C. Timbre difference capturing

Let $T_l(\boldsymbol{x})$ $(l = 1,\ldots,L)$ be the timbral metric value of attribute l for an audio sample $\boldsymbol{x}$. We estimate the timbre difference label by evaluating how much the timbral metric of the test sample deviates from the knn normal training samples. Suppose $T_l(\boldsymbol{x}^{\text{te}})$ was the r-th smallest value among $\{T_l(\boldsymbol{x}^{\text{te}}), T_l(\boldsymbol{x}_{n_1}^{\text{tr}}), \cdots, T_l(\boldsymbol{x}_{n_k}^{\text{tr}})\}$. Then, we compute the timbre difference score of $\boldsymbol{x}^{\text{te}}$ as

$$\hat{y}_l\left(\boldsymbol{x}^{\text{te}}\right) = \frac{r-1}{k} \in [0,1]. \tag{2}$$

This evaluates how large the timbral metric of the test sample is among the k-nn training samples in a nonparametric manner. Note that this value is equivalent to the special case of the U value used in the Mann-Whitney U test [21], with normalization. The Mann-Whitney U test is a nonparametric statistical test that evaluates whether the given two sets of samples are sampled from different distributions. Therefore, this value can be used for evaluating differences in the timbral metric values. Lastly, by using a predefined threshold $t \in [0,1]$, the timbre difference labels are estimated as

$$\hat{y}_l\left(\boldsymbol{x}^{\text{te}}\right) = \begin{cases} -1 & \hat{y}_l \le t \\ 0 & t < \hat{y}_l < 1-t \\ 1 & 1-t \le \hat{y}_l. \end{cases} \tag{3}$$

IV. Dataset creation

For evaluation, we created a UASD dataset with ground truth timbre difference labels. We used the MIMII DG dataset [1], a UASD dataset featuring domain generalization settings, which was also used in recent UASD competitions (DCASE Challenge Task 2 from 2022 to 2024 [14], [15], [22]). The dataset covers five machine types, each with three sections containing both source and target domain data. While MIMII DG includes machine sounds with factory noise, we also have original recordings and detailed anomaly information for each anomalous sample, enabling us to assign ground truth timbre difference labels. We refer to the original, noise-free recordings as the "clean dataset" and the MIMII DG dataset as the "noisy dataset". See [23] for further details.

The ground truth timbre difference labels for anomalous sounds were automatically generated using the timbral metrics computed by timbral models [12]. To only extract the sound difference specifically caused by machine anomaly, the ground truth labels should be determined by comparing the anomalous sounds with normal sounds that are recorded under identical conditions, including both machine operational and recording conditions. Furthermore, if the cause of the anomaly is identical, the ground truth label should also be identical. Therefore, we initially determined a single ground truth label for each condition and cause of anomalies, and then assigned these same label values to anomalous samples with identical conditions and anomaly causes.

For a single data section in the clean dataset, suppose there are M conditions and Q types of anomaly causes. Here, the conditions can be the machine's operational conditions such as machine speed, recording conditions such as microphone locations, or a combination of them. Let $\mathcal{D}_m^{\text{tr}}$ and $\mathcal{D}_{m,q}^{\text{anom}}, m = 1,\ldots,M, q = 1,\ldots,Q$, be a subset of normal training data in the clean dataset that was recorded under condition m and a subset of the anomalous data in the clean dataset for condition m and anomaly cause q, respectively. To derive the ground truth label for (m,q), we compute a score that indicates how much each timbral metric differs between these two sets of audio samples. Let $\mathcal{T}_l(\mathcal{D}) = \{T_l(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{D}\}$ denote the values of timbral metrics of timbre attribute l for a set of audio samples $\{\mathcal{D}\}$. We then evaluate the deviation of the timbral metrics between normal and anomalous samples as

$$\tilde{y}_{m,q} = \text{AUC}\left(\mathcal{T}_l\left(\mathcal{D}_m^{\text{tr}}\right), \mathcal{T}_l\left(\mathcal{D}_{m,q}^{\text{anom}}\right)\right) \in [0,1]. \tag{4}$$

Here, $\text{AUC}(\mathcal{T}_1, \mathcal{T}_2)$ denotes the area under the receiver operating characteristic curve (AUC) when $\mathcal{T}_1$ and $\mathcal{T}_2$ are regarded as the scores of negative and positive samples, respectively. Note that AUCs are also equivalent to the normalized version of the U value in the Mann-Whitney U test [21], which justifies using this score. Next, by conducting the same thresholding shown in (3) by another predefined threshold $t' \in [0,1]$, we obtain the ground truth timbre difference label $y_{m,q} \in \{1,0,-1\}$. Since t' determines from what degree of difference to notify, we tested various values ranging in $0.02 < t' < 0.2$. Finally, we assign each anomalous sample in the noisy dataset (=MIMII DG dataset) the labels computed for the corresponding condition and anomaly cause. Note that in inference, we cannot run the same procedure that created these ground truth labels, since conditions of the test data are

TABLE II: AUC of UASD (%) (for reference)

	Source						Target					
Method	Bear.	Fan	Gear.	Slider	Valve	Mean	Bear.	Fan	Gear.	Slider	Valve	Mean
Timbre-knn	62.9	64.4	55.5	68.1	63.0	62.8	49.7	45.1	52.5	54.3	55.6	51.4
Mbn-v2	**70.7**	68.5	69.4	69.0	**78.9**	**71.3**	**63.7**	47.7	**68.8**	56.2	57.7	**58.8**
PANNs	64.5	72.7	58.1	77.6	57.6	66.1	55.4	48.2	56.7	58.5	**59.5**	55.7
CLAP	65.4	69.7	66.5	**87.1**	60.2	69.8	52.8	45.4	61.3	**70.0**	58.2	57.5
BEATs	68.2	**78.8**	73.1	**81.2**	55.2	**71.3**	48.4	**54.8**	65.8	60.5	50.0	55.9

unknown and the data is corrupted with environmental noise. This is why we have to estimate the timbre difference label such as in the proposed method.

We summarize how many unique label combinations were created and the ratio of each label value in TABLE I. The variety of the ground truth labels indicates that estimating these labels can be informative in explaining the difference between normal and anomalous sounds. For example, one damage type in Bearing section 00 resulted in increased Sharpness and Boominess for a specific velocity, whereas another damage type resulted in decreased Sharpness and Brightness for the same velocity. Thus, by interviewing machine inspectors beforehand, it might be possible to automatically distinguish the anomaly causes using the acquired information.

V. Experiments

A. Experimental conditions

We conducted a joint UASD and timbre difference capturing experiment on the described dataset. While UASD results only serve as reference, **our main goal is to evaluate the timbre difference capturing performance**. For the audio encoder, we employed four models: MobileNet-v2 [24], PANNs [25], the audio encoder from CLAP [26], and BEATs [27]. MobileNet-v2 (Mbn-v2) was the baseline model in DCASE 2022 [14] and we used the same preprocessing and training procedure as in the original classification task for machine attributes. Specifically, Mbn-v2 was trained for 50 epochs using AdamW with a learning rate of 10^{-4}. For the other models, we utilized publicly available pretrained models [25]–[27] without additional modifications. Although various training methods and model architectures for the audio encoder have been explored in the literature for UASD, we chose to evaluate only the simplest approaches. Still, BEATs [27] was utilized (with further fine-tuning and model ensembling) in the top-ranking solutions in the DCASE 2024 Challenge task 2 [8], [20], so it can be seen as a simplified representative of the best current methods. For all models, we set $k = 30$. However, similar results were observed when k was set around 10 to 40. The threshold t' for the ground truth labels, and the threshold t for each timbre difference capturing method were both set to multiple values $\{0.02, 0.05, 0.1, 0.15, 0.20\}$, and the average performance over all the values was computed.

For comparison, there are no existing methods in the literature that can be performed for this task. This is because existing audio difference captioning methods [9], [10] require normal-anomaly pairs of audio data and captions for training, whereas no anomalous data are available for training in UASD. Additionally, the existing timbre-based ASD method [19] was also a supervised ASD method, thus cannot be performed in our problem setting. For this reason, we designed and performed three baseline methods to evaluate the effectiveness of the proposed method. The first method uses a constant timbre difference label as output, $\hat{y}_l = 0$ (Constant). The second method estimated the labels by comparing the anomalous sample's timbral metrics with all normal training samples using (2) (Compare-All). This served as the baseline for timbre difference label estimation without normal sample selection. In the third method, we performed UASD and timbre difference capturing with the L timbral metrics as the features for knn, instead of the audio embeddings (Timbre-knn). Comparing with this method allowed us to evaluate the effectiveness of using audio encoder embeddings.

The UASD performance was evaluated by AUC. The timbre difference capturing performance was evaluated by the mean absolute error (MAE) for anomalous samples, where we normalized the error to compensate for imbalances in ground truth labels [28]. That is, for timbre l, the MAE value is

$$\mathrm{MAE}_l = \frac{1}{3}\sum_{i=1}^{M} \frac{|\hat{y}_l\left(\boldsymbol{x}_i^{\mathrm{te}}\right) - y_{l,i}|}{M_{l,y_{l,i}}}, \tag{5}$$

where $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{M}$ are the anomalous test data, $y_{l,i}$ is the ground truth timbre difference label for timbre l of sample $\boldsymbol{x}_i$, $M_{l,y} = |\{y_{l,i} \mid y_{l,i} = y, 1 \le i \le M\}|$ is the number of samples in the test data that the ground truth label value is identical to y.

B. Results

For reference, we show the AUC values of each method in TABLE II. All four audio encoders had higher average AUC values than Timbre-knn. This indicates that timbral metrics alone are not good features for detecting anomalies, whereas the audio embedding spaces obtained from the models can better distinguish anomalies. Still, describing differences with timbre for each machine condition can be useful, which is pursued by the proposed method (Fig. 1).

Next, Table III shows the MAE of timbre difference capturing, which is the main result of our experiment. In the source domain, PANNs, CLAP, and BEATs had smaller average MAE than the three baseline methods, indicating the effectiveness of our proposed strategy of comparing anomalous samples with nearby normal samples. Notably, BEATs yielded a much lower MAE than the baseline methods, consistent with its highest AUC score. BEATs performed particularly well on Bearing, Gearbox, and Slider, which were the machines that had many conditions that can alter their sounds (see "#g" in Table I). This indicates that the proposed method achieves its initial goal of accurately estimating timbre differences even when normal sounds vary widely.

By contrast, although MobileNet-v2 resulted in the same average AUC value as BEATs, its average MAE was mostly the same as the baseline methods. A possible explanation is that MobileNet-v2 was only trained on normal sounds from the target machines, which limited its ability to assess how similar the normal sounds are to out-of-distribution anomalous data—even though it could recognize when a sample was out-of-distribution. In contrast, the other audio encoders were trained on a broad spectrum of sound data, likely producing

TABLE III: MAE of timbre difference capturing (↓). Average and standard deviations across threshold values t are shown.

	Source						Target					
Method	Bearing	Fan	Gearbox	Slider	Valve	Mean	Bearing	Fan	Gearbox	Slider	Valve	Mean
Constant	0.56 ± 0.06	0.61 ± 0.10	0.74 ± 0.04	0.57 ± 0.06	0.46 ± 0.13	0.59	0.67 ± 0.06	0.66 ± 0.10	0.71 ± 0.05	0.69 ± 0.06	0.67 ± 0.05	0.68
Compare-All	0.45 ± 0.06	0.57 ± 0.10	0.66 ± 0.04	0.61 ± 0.07	0.49 ± 0.16	0.56	**0.56 ± 0.05**	0.64 ± 0.09	0.68 ± 0.04	0.65 ± 0.04	**0.62 ± 0.05**	0.63
Timbre-knn	0.46 ± 0.05	0.55 ± 0.07	0.71 ± 0.04	0.57 ± 0.07	0.51 ± 0.15	0.56	0.59 ± 0.08	0.64 ± 0.09	0.68 ± 0.05	0.74 ± 0.08	0.65 ± 0.06	0.66
Mbn-v2 (Proposed)	0.44 ± 0.05	0.57 ± 0.07	0.77 ± 0.04	0.58 ± 0.06	0.51 ± 0.14	0.58	0.60 ± 0.07	**0.61 ± 0.07**	0.70 ± 0.04	0.63 ± 0.05	0.59 ± 0.06	0.63
PANNs (Proposed)	0.45 ± 0.06	0.52 ± 0.06	0.61 ± 0.04	0.50 ± 0.08	0.49 ± 0.15	0.52	0.58 ± 0.06	0.64 ± 0.08	0.67 ± 0.05	0.61 ± 0.03	0.68 ± 0.09	0.64
CLAP (Proposed)	0.42 ± 0.05	0.52 ± 0.06	0.61 ± 0.04	**0.48 ± 0.07**	0.48 ± 0.15	0.50	0.57 ± 0.07	0.63 ± 0.07	**0.66 ± 0.05**	**0.58 ± 0.04**	0.64 ± 0.08	**0.62**
BEATs (Proposed)	**0.35 ± 0.05**	**0.50 ± 0.04**	**0.58 ± 0.04**	**0.48 ± 0.07**	**0.45 ± 0.15**	**0.47**	0.60 ± 0.07	**0.61 ± 0.07**	0.67 ± 0.05	**0.62 ± 0.06**	0.62 ± 0.08	**0.62**

embedding spaces that better capture similarities between normal and anomalous machine sounds. In the target domain, the various methods all showed similar MAE values, possibly because of the limited dataset size. Addressing this issue lies beyond this paper's scope and future work to solve it will be needed. Overall, these findings suggest that by using audio encoders with rich embedding spaces, the proposed method can accurately estimate timbre differences even under significant variation in normal data, provided that sufficient training data are available.

VI. Conclusion

We proposed a framework to explain anomaly differences based on pre-selected timbral attributes, which does not require anomalous data or ground truth labels for training. We then further proposed a timbre difference capturing method that works alongside UASD, comparing anomalous test samples only with their most similar normal training examples. We achieved this by using audio encoders and k-nearest neighbors in the embedding space the encoders generate. Experiments using the MIMII DG-based dataset, with additionally generated ground truth timbre difference labels, confirmed that the proposed method can accurately estimate the timbre difference labels even when the normal training sounds have a variety due to various conditions.

VII. Acknowledgment

We would like to express our sincere gratitude to Prof. Keisuke Imoto of Doshisha University for his invaluable advice throughout this research.

References

[1] Y. Koizumi *et al.*, "Description and discussion on DCASE2020 Challenge Task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2020, pp. 81–85.

[2] 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 ICASSP*, May 2020, pp. 271–275.

[3] 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," DCASE Challenge, Tech. Rep., 2020.

[4] 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.

[5] J. Lopez, G. Stemmer, and P. Lopez-Meyer, "Ensemble of complementary anomaly detectors under domain shifted conditions," DCASE Challenge, Tech. Rep., 2021.

[6] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?" *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 608–622, 2024.

[7] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *Proc. IEEE ICASSP*, 2024, pp. 276–280.

[8] Z. Lv *et al.*, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE Challenge, Tech. Rep., June 2024.

[9] S. Tsubaki *et al.*, "Audio-change captioning to explain machine-sound anomalies," in *Proc. DCASE Workshop*, September 2023, pp. 201–205.

[10] D. Takeuchi, Y. Ohishi, D. Niizumi, N. Harada, and K. Kashino, "Audio difference captioning utilizing similarity-discrepancy disentanglement," in *Proc. DCASE Workshop*, September 2023, pp. 181–185.

[11] K. Jensen, "The timbre model," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2238–2238, 2002.

[12] A. Pearce, T. Brookes, and R. Mason, "Timbral attributes for sound effect library searching," in *J. Audio Eng. Soc.*, 2017.

[13] Y. Kawaguchi *et al.*, "Description and discussion on DCASE 2021 Challenge Task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE Workshop*, November 2021, pp. 186–190.

[14] K. Dohi *et al.*, "Description and discussion on DCASE 2022 Challenge Task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE Workshop*, November 2022.

[15] ——, "Description and discussion on DCASE 2023 Challenge Task 2: First-Shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, September 2023, pp. 31–35.

[16] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[17] K. Minemura, T. Ogawa, and T. Kobayashi, "Acoustic feature representation based on timbre for fault detection of rotary machines," in *2018 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*. IEEE, 2018, pp. 302–305.

[18] T. Mian, A. Choudhary, and S. Fatima, "An efficient diagnosis approach for bearing faults using sound quality metrics," *Applied Acoustics*, vol. 195, p. 108839, 2022.

[19] Y. Ota and M. Unoki, "Anomalous sound detection for industrial machines using acoustical features related to timbral metrics," *IEEE Access*, vol. 11, pp. 70 884–70 897, 2023.

[20] A. Jiang *et al.*, "Thuee system for first-shot unsupervised anomalous sound detection," DCASE Challenge, Tech. Rep., June 2024.

[21] S. J. Mason and N. E. Graham, "Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation," *Q. J. R. Meteorol. Soc.*, vol. 128, no. 584, pp. 2145–2166, 2002.

[22] T. Nishida *et al.*, "Description and discussion on dcase 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," 2024. [Online]. Available: https://arxiv.org/abs/2406.07250

[23] K. Dohi *et al.*, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proc. DCASE Workshop*, 2022.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR)*, 2018, pp. 4510–4520.

[25] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[26] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. IEEE ICASSP*, 2024, pp. 336–340.

[27] S. Chen *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, vol. 202, July 2023, pp. 5178–5193.

[28] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *Proc. Int. Conf. Intel. System. Des. and Appl.* IEEE, 2009, pp. 283–287.