Retrieval-Augmented Difference Captioning to Explain Unsupervised Anomalous Sound Detection

Ryoya Ogura, Tomoya Nishida, Yohei Kawaguchi
Research and Development Group, Hitachi, Ltd.
{ryoya.ogura.bx, tomoya.nishida.ax, yohei.kawaguchi.xk}@hitachi.com

***Abstract*—This paper proposes a method for generating explanatory captions that describe how sounds identified as anomalous in unsupervised anomalous sound detection (UASD) differ from normal sounds. In previous methods, anomaly detection and caption generation are performed independently, so captions may not reflect the reasons for detection. Furthermore, it requires paired normal and anomalous sounds for training, making it impossible to apply to UASD, where anomalous sounds are not available for training. To address these challenges, the proposed method utilizes a feature from the pre-trained CLAP (Contrastive Language-Audio Pre-training) model for both UASD and caption generation, ensuring consistency between detection and explanation, and enabling caption generation without training. For sounds identified as anomalous, we use retrieval-augmented generation (RAG) to search the training data for similar normal sounds and compare captions. RAG enables the capture of the differences between anomalous and normal sounds. Experiments based on subjective evaluation and a sample-wise analysis of the output captions demonstrate the effectiveness of the proposed method.**

***Index Terms*—anomalous sound detection, captioning, CLAP, RAG**

I. Introduction

Unsupervised anomalous sound detection (UASD) is the task of identifying whether sounds are normal or anomalous by only using normal sounds as training data, which is applied to inspection using operational sounds of industrial machines [1–3]. Automatic detection of mechanical failure is essential for artificial intelligence (AI)–based factory automation, as timely detection of machine anomalies via sound is an effective method for machine condition monitoring.

In UASD, deep learning models, such as autoencoders [4], are used. In recent years, the use of large-scale pre-trained models [5] and learning with artificially generated labels [6] have been reported to show high performance. On the other hand, for machine inspection, it is important not only to detect anomalous sounds but also to help identify their causes. If it is possible to explain how anomalous sounds differ from normal sounds, it makes it easier to identify the cause of the anomaly.

To analyze anomalous sounds, Tsubaki et al. [7] proposed a method that inputs both normal and anomalous sounds and outputs captions describing their differences. We will refer to these captions as "difference captions." In their method, pairs of normal and anomalous sounds are annotated with their differences to create the MIMII-Change dataset, which is used to train the model. The model's audio encoder outputs normal and anomalous sound embeddings. These embeddings and the subtraction of these embeddings are concatenated and passed to a transformer-based decoder, which generates difference captions. Recently, such methods for learning the differences between two audio pairs have also been studied for targets other than mechanical sounds [8], [9].

However, since the caption generation method proposed by Tsubaki et al. operates independently from the anomaly detection process. As a result, the generated captions may sometimes be unrelated to the outcomes of anomaly detection. Furthermore, in UASD, methods that rely on paired normal and anomalous sounds for training cannot be used, because such pairs do not exist for training data. Instead, we must explain how an unseen anomalous sound deviates from the learned distribution of normal sounds without having anomalous sounds for training. Therefore, it is not possible to apply existing difference captions generation methods [7–9] to UASD.

In this paper, we propose a difference captioning method that suits the UASD problem setting. Our proposed method uses embeddings from an audio-language model for both UASD and difference captions generation. By performing both UASD and caption generation in the same joint audio–language embedding space, the detection results and the generated captions are expected to be less likely to contradict each other. In this method, only normal sounds in the training data for UASD are required, and caption generation can be performed without training. The difference captions are generated based on retrieval-augmented generation (RAG) [10]. First, a caption is generated from the embedding of each test sample identified as anomalous by the trained UASD system using a pre-trained audio-language model's text decoder. Second, normal sounds most similar to each test sample sound are retrieved from the training data for UASD, and captions of these normal sounds are generated by the text decoder. Finally, by comparing the captions of each test sample that is identified as anomalous and normal sound captions with GPT-4 [11], a difference caption is generated. With the introduction of RAG and pre-trained text decoder, our proposed method can explain how anomalous sounds differ from normal sounds without learning anomalous sounds themselves, making it suitable for UASD. We employ CLAP (Contrastive Language-Audio Pre-training) [12] as a pre-trained audio-language model. In our experiments conducted under the UASD problem setting, we first found that CLAP shows sufficient performance on anomalous sound detection. More importantly, by observing the different captions for the sounds identified as anomalous, we were able to

979-8-3315-7206-8/25/$31.00

obtain reasonably valid captions that matched the cause of the anomalies.

II. Related Work

A. CLAP

CLAP is a foundation model that generates embeddings, retaining both audio and text information by mapping audio and text into a joint multimodal space through contrastive learning. Its audio and text encoders are trained to maximize similarity for matching audio–text pairs while minimizing similarity for non-matching pairs. Utilizing CLAP facilitates zero-shot classification: by calculating the similarity between an input audio's embedding and the text embeddings of various labels, the most appropriate label naturally exhibits the highest similarity. This enables sound event classification without requiring pre-labeled audio for training.

The experiments in this paper utilize CLAP, which was released by Microsoft in 2023 [12]. This CLAP has been pre-trained on a large scale using datasets such as WavCaps [13] and AudioSet [14]. Audio transformers (HTSAT [15]) are used for the audio encoder, and GPT-2 [16] is used for the text encoder within CLAP. Additionally, the text decoder, consisting of a mapper network and GPT-2, can generate sentences from an embedding, which can be combined with the audio encoder to perform audio captioning.

B. Retrieval-Augmented Audio Caption (RECAP)

RECAP is a method proposed by Ghosh et al. [17] for audio captioning. Audio captioning is the task of describing input environmental sounds in text, and most previous models consist of a pre-trained audio encoder and a text decoder. However, these encoder-decoder architectures do not work well when the input data is from a domain different from the training data. To address the problem caused by this domain shift, Ghosh uses CLAP to measure the similarity between the input sound and each caption in the data store. The top four similarity captions are input to GPT-2 to estimate the caption of the input sound.

RECAP is based on RAG, where information from an external data store is added to the training data for generation. We considered RAG to be effective in captioning anomalous sounds. For example, as pre-trained models such as CLAP are not necessarily trained by the sounds of industrial equipment, generating captions of anomalous machine sounds may not result in captions explaining how the sound is anomalous. However, adding additional information, such as captions for normal sounds, allows comparisons between normal and anomalous sounds and can create detailed captions for anomalous sounds.

III. Proposed Method

A. UASD with CLAP's audio encoder

Figure 1 shows the overview of the proposed method. The training dataset, comprising solely normal sounds, and the test dataset, encompassing both normal and anomalous sounds, are fed into CLAP's audio encoder to derive corresponding embeddings for each dataset. For anomaly detection, k-Nearest Neighbors (k-NN) based on Euclidean distance is applied. The anomaly score $\mathcal{A}(\mathbf{x})$ is defined as

$$\mathcal{A}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k}\|\mathbf{x} - \mathbf{X}_i\|_2, \tag{1}$$

where $\mathbf{x}$ is the embedding of a test sample. $\mathbf{X}_i$ $(i = 1, \ldots, k)$ are the embeddings of the reference samples from the training database, e.g., $\mathbf{X}_i$ is the i-th closest embedding in the training database to x. Anomalous sounds tend to diverge from the distribution of normal sounds, resulting in higher anomaly scores compared to normal sounds in the test dataset. The test samples are identified based on a threshold: those with anomaly scores above the threshold are classified as anomalous, while those with scores below or equal to the threshold are considered normal.

B. Difference caption generation for anomalous sounds

The test samples whose anomaly scores exceeded a certain threshold were identified as anomalous, and captions were generated for them. To optimize the threshold, we set it to the point on the receiver operating characteristic (ROC) curve computed on the test dataset that is closest, in Euclidean distance, to the top-left corner (0, 1). The basic idea to generate difference captions without training is to first independently create captions for each anomalous and k-nearest normal sounds using a pre-trained text decoder, and then let large language models, GPT-4 in our experiments, compare those output captions. The approach of comparing captions is inspired by RAG. In our method, k-NN not only serves for anomaly detection in UASD, but also plays the role of retrieving the most similar normal sounds from the reference samples for each test sample. Instead of generating captions solely from the anomalous sounds, providing the captions of the most similar normal sounds as external data and prompting a comparison enables the model to more effectively capture subtle differences between anomalous and normal sounds.

To create captions of each sound, we utilize the CLAP embeddings used in UASD. The embedding of the sound identified as anomalous in UASD, $X^{(a)}$, is input to the CLAP's text decoder to generate a caption that explains this sound. Embeddings of the k nearest samples $X_1, \ldots X_k$ are also fed into the same text decoder to generate captions for k instances of normal sounds. All these $k + 1$ captions are then used to form a prompt that asks GPT-4 how the caption of the anomaly-identified sound is different from the other captions. Specifically, the prompt is formed as *"The caption of the anomalous sound of [machine name] is given as: [caption of the anomaly-identified sound]. On the other hand, the captions of the normal sounds of [machine name] are given as: [k captions of the normal sounds, separated with commas]. Please describe in broad strokes how this anomalous sound differs compared to the normal sounds."* The output of GPT-4

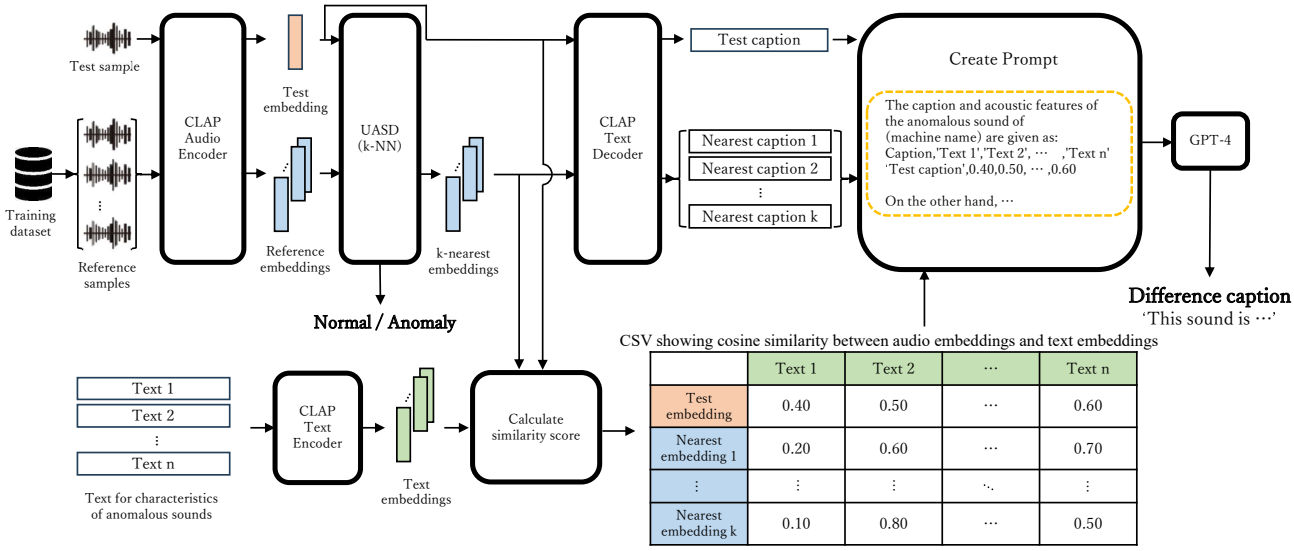

Fig. 1. UASD and captioning flow in our proposed method.

is then directly used as the difference caption. In this paper, we will call this method the "Text decoder-based method."

This method performs both UASD and difference caption generation in the same CLAP embedding space. It is expected that this ensures consistency between the UASD results and the content of the difference captions. Appropriate difference captions are expected to be obtained for cases where the anomalous and normal sounds have differences that CLAP's captioning network can capture. We assume that such a case holds for some typical anomalous conditions since anomalous conditions can often cause additional sounds with different characteristics than the original normal sounds, e.g., rattling sounds added to smooth movement sounds, or cause certain sounds to disappear, which is likely to appear in the captions.

C. Difference caption generation with predefined texts

While certain differences between the anomalous and normal sounds can be captured by the first method, in some cases, CLAP's text decoder can provide very similar descriptions for the anomalous sound and the k reference normal sounds to be compared. This can happen even when the distance between the embeddings is large and the anomaly score is high. In such cases, comparing those descriptions in GPT-4 cannot acquire appropriate captions that explain the differences between anomalous and normal sounds.

To solve this problem, we additionally propose a method based on the zero-shot classification framework using CLAP. First, a set of L short reference texts $text_1, \ldots, text_L$, describing common characteristics of malfunctioning machine sounds, such as "Vibration" and "Popping and Knocking sounds" are prepared. These texts can for example be prepared by asking large language models. For each of these texts, the text embeddings T_l $(l = 1, \ldots, L)$ are extracted through CLAP's text encoder. Then, the cosine similarities $S_C(X, T_l)$ between each text embedding T_l and audio embedding $X \in \{X^{(a)}, X_1, \ldots, X_L\}$ are computed to infer how much that audio input contains the feature described in the lth text. Finally, the obtained similarity scores are included into a prompt that asks to compare those values for the anomaly-identified sound and the normal reference sounds. Specifically, the prompt is formed as *"The acoustic features of the anomalous sound of [machine name] are given as [similarity scores for the anomaly-identified sound]. On the other hand, the acoustic features of the k normal sounds of [machine name] are given as [similarity scores for the k reference normal sounds]"*. Here, the similarity scores are given as a CSV format where the rows represent scores for each reference text and the columns represent scores for each sound, e.g., for the reference normal sounds, the scores are given as

$$
\begin{aligned}
&text_1,\ text_2,\ \ldots,\ text_L,\ \backslash\mathrm{n}\\
&S_C(X_1, T_1),\ S_C(X_1, T_2),\ \ldots,\ S_C(X_1, T_L),\ \backslash\mathrm{n}\\
&\ldots,\\
&S_C(X_k, T_1),\ S_C(X_k, T_2),\ \ldots,\ S_C(X_k, T_L),
\end{aligned}
$$

if the similarity between some reference text and the sound is different for the normal sounds and the anomaly-identified sound, that difference should be extracted as the final output caption. We will call this method the "Zero-shot classification-based method."

This strategy can also be combined with the method proposed in the previous section. This is realized by including both the caption information and the similarity scores in one CSV format, such as

$$
\begin{aligned}
&\text{caption},\ text_1,\ text_2,\ \ldots,\ text_L,\ \backslash\mathrm{n}\\
&\text{[output caption]},\ S_C(X_1, T_1),\ S_C(X_1, T_2),\ \ldots,\ S_C(X_1, T_L),\ \backslash\mathrm{n}\\
&\ldots,
\end{aligned}
$$

in this way, the output difference caption would include both caption based and specified characteristics-based explanations.

TABLE I
AUC(%) FOR BASELINE AND PRE-TRAINED MODELS.

Machine	ID	Autoencoder	pre-trained models		
			PANNs	LAION-CLAP	**MS-CLAP (Proposed)**
ToyCar	01	**81.36**	80.04	76.94	75.68
	02	**85.97**	85.92	84.11	80.99
	03	63.30	**69.98**	64.76	65.42
	04	84.45	**89.42**	85.53	83.57
	Average	78.77	**81.34**	77.84	76.42
ToyConveyor	01	**78.07**	61.65	63.14	63.42
	02	**64.16**	56.45	55.78	59.23
	03	**75.35**	60.24	58.28	59.65
	Average	**72.53**	59.45	59.07	60.77
Fan	00	**54.41**	50.63	53.87	51.91
	02	**73.40**	60.25	65.21	61.89
	04	**61.61**	45.77	47.78	46.88
	06	73.92	**75.21**	74.40	67.96
	Average	**65.83**	57.97	60.32	57.16
Pump	00	67.15	81.92	**87.40**	84.22
	02	61.53	68.23	89.13	**92.49**
	04	**88.33**	77.71	83.27	80.39
	06	**74.55**	56.83	62.78	58.83
	Average	72.89	71.17	**80.64**	78.98
Slider	00	96.19	99.89	**99.98**	99.25
	02	78.97	88.58	91.25	**93.87**
	04	**94.30**	86.51	79.21	80.46
	06	69.59	57.31	61.99	**71.28**
	Average	84.76	83.07	83.11	**86.22**
Valve	00	68.76	84.38	**91.07**	80.53
	02	68.18	**79.89**	79.39	76.61
	04	74.30	79.26	80.80	**81.45**
	06	53.90	69.53	64.22	**70.23**
	Average	66.28	78.26	**78.87**	77.21
Overall average		**73.51**	71.88	73.30	72.79

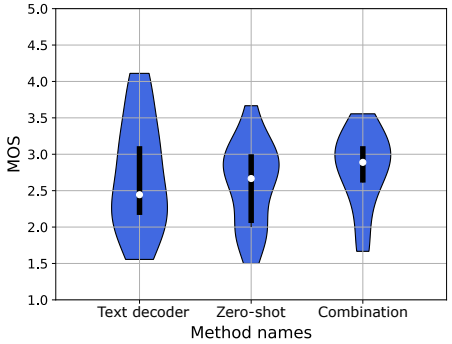

Fig. 2. Violin plot for MOS values of each ID in each machine. "Text decoder", "Zero-shot", and "Combination" denote Text decoder-based method, Zero-shot classification-based method, and combination of these methods, respectively. White dot denotes the median and black bar denotes range of quartiles.

IV. EXPERIMENT

A. Experimental conditions

To verify the effectiveness of the proposed method under simple conditions where no domain shifts occur, such as changes in operational conditions, we used the DCASE 2020 Challenge Task 2 Development Dataset [4] for the experiments. It contains four types of machines (Fan, Pump, Slider, Valve) from MIMII Dataset [18] and two types of toys (ToyCar, ToyConveyor) from ToyADMOS [19] with normal and anomalous sounds. Each of the four machines from the MIMII Dataset has four Machine IDs (00, 02, 04, 06). ToyCar also has four IDs (01, 02, 03, 04), and ToyConveyor has three IDs (01, 02, 03). ID is the identifier of each individual of the same machine type.

In this experiment, we set $k = 4$. The performance of UASD per Machine ID is calculated using the area under the ROC (AUC). Note that the specific anomaly score threshold we selected for caption generation does not influence the AUC, as AUC considers all possible thresholds. The CLAP's audio encoder by Microsoft used in our proposed method was compared in performance with pre-trained models such as PANNs [20] and the CLAP audio encoder by LAION [21], in addition to the autoencoder used as a baseline for the DCASE 2020 Challenge Task 2 [4].

Next, captions were generated for the test data identified as anomalous sounds in UASD. Three methods of generation were used: the text decoder-based method, the zero-shot classification-based method, and the combination of both methods. For all methods, a common prompt instructing the GPT-4 to begin their output sentence with "This sound is" and to finish it within 40 words was added before each prompt described in the previous section. For the text decoder-based method and the combined method, we constrained the text output from CLAP's text decoder to be in the form of "Sounds like ...". We asked GPT-4 to generate eight descriptions of anomalous sounds for the zero-shot classification-based method: "Vibration", "High-frequency Squealing or Screeching", "Popping or Knocking Sounds", "Rhythmic Clicking or Tapping", "Grinding Sounds", "Irregular Patterns", "Low-frequency Humming", and "Unexpected Silence".

We subjectively evaluated the output captions by the Mean Opinion Score (MOS). For evaluation, we selected three samples for each ID of each machine type that showed the highest anomaly scores, since captions for high anomaly score samples matter the most. In total, 69 sets of data-caption pairs were evaluated for each method. Three non-expert subjects were asked to listen to the test and reference normal samples, and rate how well each caption explains the differences between them. The ratings ranged from '1' to '5', where '1' represents the worst and '5' represents the best. In addition, we reviewed each created caption to verify if they align with the cause of the anomaly and the sound changes that should occur from that anomaly.

B. UASD experimental results

Table I shows the experimental results of UASD using CLAP embeddings along with the baseline and other pre-trained models. "MS-CLAP" refers to the CLAP by Microsoft. The results indicate that the MS-CLAP's embedding achieves a performance level almost parallel to that of other pre-trained models, with the performance variance between CLAP and the other pre-trained models dependent on the specific model and Machine ID. This discrepancy in performance may be attributed to the slight differences in the dataset used for pre-training the models. The findings also suggest that by applying transfer learning and fine-tuning to CLAP's audio encoder, UASD performance could potentially be enhanced

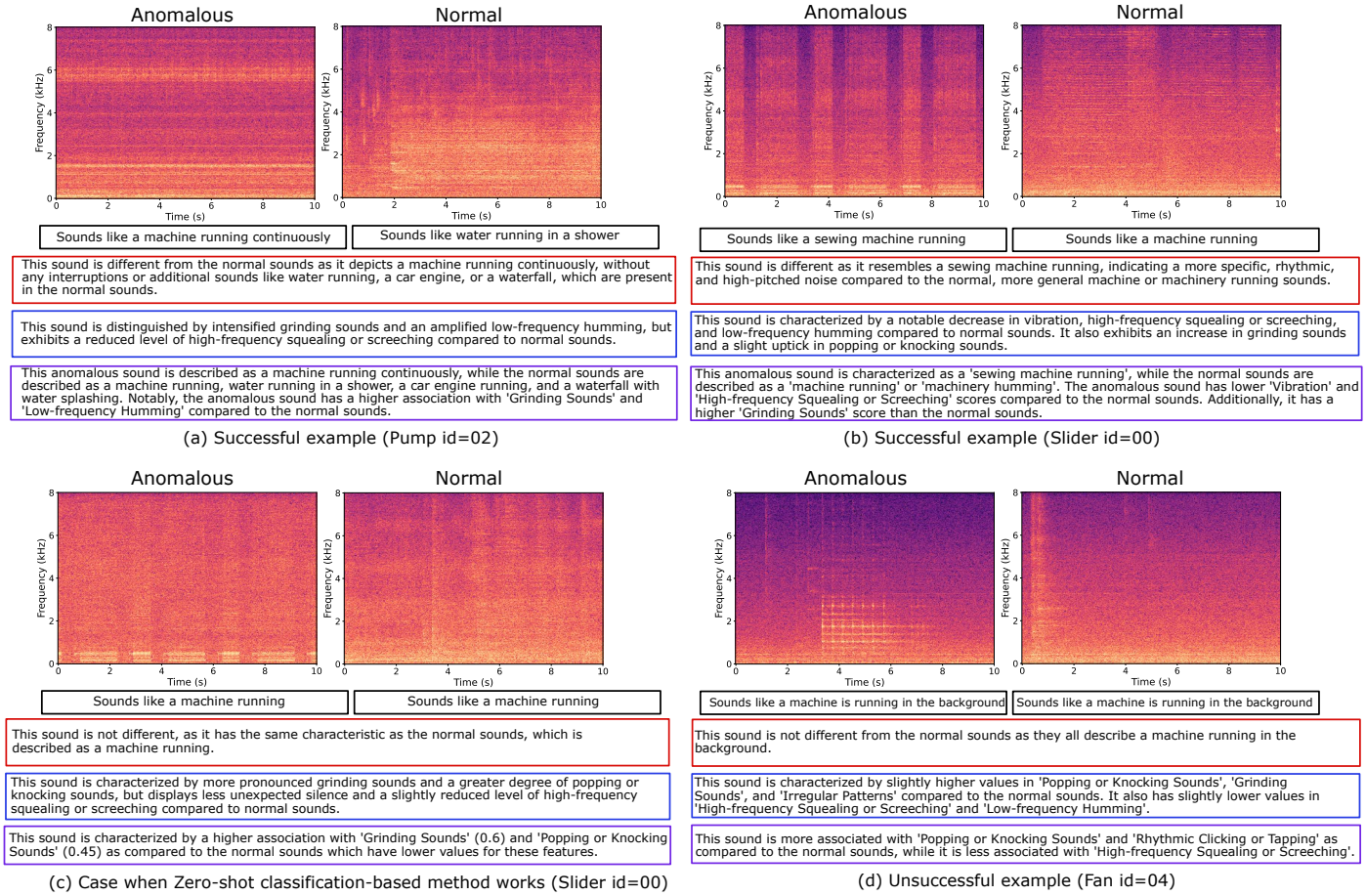

Fig. 3. Spectrograms of anomalous and reference normal sounds and corresponding captions. Black box (first row) denotes individual captions for each sample; Red box (second row) denotes captions for CLAP's text decoder-based method; Blue box (third row) denotes captions for zero-shot classification-based method; Purple box (fourth row) denotes captions for combination of both methods.

further, leveraging the pre-trained model's capabilities more effectively.

MS-CLAP underperforms the baseline with the autoencoder in machine types of toy-car, toy-conveyor, and fan. On the other hand, the average of all machine types' AUC values is 73.51% for the baseline and 72.79% for MS-CLAP, so the overall performance is similar.

C. Captioning experimental results

Figure 2 shows the distribution of the MOS values computed for each ID in each machine. Overall, the MOS values for a considerable portion of IDs were over '3', indicating that the output captions can be informative to some extent. The number of IDs that received MOS values over '3' was 10, 8, and 11 out of 23, for the text-decoder-based method, the zero-shot classification-based method, and the combination method, respectively. The text-decoder-based method showed significantly higher MOS values for several IDs than other methods, such as Pump ID 00 (4.1), Slider ID 00 (4.0), and ToyCar ID 01 (3.8), indicating its usefulness for certain kinds of data. At the same time, some IDs exhibited relatively low MOS values between 1.5 to 2.5, and text decoder-based method showed more variation in MOS across IDs than the other methods. This was mainly because the text decoder sometimes provided very similar captions for both anomalous and normal sounds, resulting in difference captions that are not very expressive. On average, the combination of the two proposed methods showed the highest MOS value, which indicates the combination method was able to take the good parts of both methods to some extent.

Figure 3 shows examples of the difference captions generated by the three methods against anomalous test samples. We picked typical examples to show the general trends of the proposed methods. (a) shows an example for pumps, where the text decoder-based method describes the absence of water sounds in the anomalous sound. This coincides with the cause of malfunction in pumps, where pumps in normal conditions have water flows but anomalous condition does not, due to leakage or clogging [18]. The zero-shot classification-based method generated the difference caption focusing on the low frequency of the anomalous sound, which can also be seen in the spectrograms. (b) shows an example for sliders, which is also a successful example. In this case, an anomalous slider makes sounds caused by scrapings or rattlings of rails due

to rail damage or lack of grease [18]. Expressions such as "sewing machine" and "grinding" match such characteristics, explaining strong periodic sounds or sounds related to hard substances scraping each other. (c) is another slider example, where the text decoder-based method did not capture any differences. However, the zero-shot classification-based method successfully described the difference, explaining the difference in "grinding" sounds. This shows the effectiveness of the zero-shot classification-based method, even when the text decoder-based method does not work ideally. (d) is an example for fans, where none of the methods successfully described the difference. While the zero-shot classification-based method described differences such as higher "Popping or Knocking sounds" or "Irregular Patterns", they seem to just coincide with the background noise that appears in 3.5(s)–6 (s) in the anomalous sound. Considering the fan's low AUC values, it can be inferred that obtaining appropriate difference captions for machine types with low AUC values is difficult.

V. Conclusion

In this paper, we proposed a method for jointly conducting UASD and difference captioning without paired data of normal and anomalous sounds for training based on CLAP embedding. The use of CLAP embedding confirmed that UASD is possible with high performance. In addition, results indicated that the RAG-based method for difference-caption generation can generate captions that match why the samples are detected as anomalous to some extent.

VI. Acknowledgment

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

References

[1] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv*, 2025.

[2] 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, 2019.

[3] 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*, 2020, pp. 271–275.

[4] 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. DCASE Workshop*, 2020, pp. 81–85.

[5] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W.-Q. Zhang, P. Fan, J. Liu, and Y. Qian, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *Proc. IEEE ICASSP*, 2024, pp. 1326–1330.

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

[7] S. Tsubaki, Y. Kawaguchi, T. Nishida, K. Imoto, Y. Okamoto, K. Dohi, and T. Endo, "Audio-change captioning to explain machine-sound anomalies," in *Proc. DCASE Workshop*, 2023, pp. 201–205.

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

[9] S. Deshmukh, S. Han, R. Singh, and B. Raj, "ADIFF: Explaining audio difference using natural language," in *ICLR*, 2025.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, 2020, pp. 9459–9474.

[11] OpenAI, J. Achiam, *et al.*, "GPT-4 technical report," *arXiv:2303.08774*, 2023.

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

[13] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3339–3354, 2024.

[14] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, 2017, pp. 776–780.

[15] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. IEEE ICASSP*, 2022, pp. 646–650.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *Open AI blog*, 2019.

[17] S. Ghosh, S. Kumar, C. K. Reddy Evuru, R. Duraiswami, and D. Manocha, "Recap: Retrieval-augmented audio captioning," in *Proc. IEEE ICASSP*, 2024, pp. 1161–1165.

[18] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proc. DCASE Workshop*, 2019, pp. 209–213.

[19] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proc. IEEE WASPAA*, 2019, pp. 308–312.

[20] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[21] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE ICASSP*, 2023.