Hitachi Review

January 2025 Technical Information

Text Generation Technology for Explaining Changes in Sound Emitted by Machines Operating Abnormally

#Productivity Improvement #Generative AI #IoT/Data Utilization #Research & Development

Author

Yohei Kawaguchi, Ph.D. Eng.
Media Intelligent Processing Research Department, Advanced Artificial Intelligence Innovation Center, Center for Digital Services, Research & Development Group, Hitachi, Ltd.
Current work and research: Management of research and development of foundation models and AI for audio, acoustic and time-series signal processing.
Society memberships: Senior member of IEEE, member of the Acoustical Society of Japan, the Institute of Electronics, Information and Communication Engineers (IEICE), and the Information Processing Society of Japan (IPSJ)

Tomoya Nishida
Media Intelligent Processing Research Department, Advanced Artificial Intelligence Innovation Center, Center for Digital Services, Research & Development Group, Hitachi, Ltd.
Current work and research: Research and development of acoustic signal processing AI and foundation models.
Society memberships: Acoustical Society of Japan

Kota Dohi
Media Intelligent Processing Research Department, Advanced Artificial Intelligence Innovation Center, Center for Digital Services, Research & Development Group, Hitachi, Ltd.
Current work and research: Research and development of time-series signal processing AI and foundation models.
Society memberships: Acoustical Society of Japan

Takashi Endo
Media Intelligent Processing Research Department, Advanced Artificial Intelligence Innovation Center, Center for Digital Services, Research & Development Group, Hitachi, Ltd.
Current work and research: Research and development of acoustic signal processing AI and foundation models.
Society memberships: Acoustical Society of Japan, IEICE

Highlight

Listening checks that rely on the subjective impression of an experienced technician have been a common practice in many different industries, including quality inspections on production lines and field inspections of infrastructure. Unfortunately, these in-person checks are becoming less practical as the shrinking of the workforce over recent years is causing a shortage of personnel with the necessary skills.

In response, Hitachi has been working to develop technology and offer solutions for automating sound-based inspection practices and making them more efficient. Past practice has been to input the sound of machine operation into a system that outputs a result indicating the extent to which it diverges from the normal sound, or that issues an alert if this divergence exceeds a threshold. This article describes an AI technique that, instead of just using data on machine operating sounds as a basis for issuing alerts, also generates text explaining the detected anomaly in a way that prompts maintenance actions.

1. Introduction

Listening checks that rely on the subjective impression of an experienced technician have been a common practice in many different industries, including quality inspections on production lines and field inspections of infrastructure. Unfortunately, these in-person checks are becoming less practical as the shrinking of the workforce over recent years is causing a shortage of personnel with the necessary skills. In response, Hitachi has been working on developing and supplying solutions for automating sound-based inspection[1), 2)] and on an AI technique for anomalous sound detection that forms the core of these solutions. This technology development has involved open innovation in the form of publishing data sets containing the sounds of operating industrial equipment[3) - 5)] and

sponsoring international competitions for anomalous sound detection[6) - 10)]. It has also included the development of various in-house artificial intelligence (AI) techniques[11) - 22)]. Through this work, Hitachi believes it has considerably expanded the scope of practical application for anomalous sound detection.

Past anomalous sound detection techniques have used the sound of machine operation as an input and have output a result indicating the extent to which the input sound diverges from the normal sound, or have issued an alert if this divergence exceeds a threshold[23)]. It was felt, however, that if an AI provided information on why a sound was judged to be anomalous rather than just issuing an alert, it could indicate the maintenance actions to take in response.

This article describes a technique for acoustic change captioning that generates text explaining how sound characteristics have changed. It works by having an AI compare any abnormal sounds that are detected with a database of prerecorded normal sounds. The technique was presented at the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) 2023[24)] and more details may be found in the published paper.

2. Past Research

Image change captioning has been a subject of past research. This technique involves the automatic generation of text explaining the changes between two (before and after) input images. In 2018, a method based on pixel difference was proposed to pinpoint the areas of change between images. While this method was able to identify significant scene changes, it could not effectively distinguish between important changes and variations in viewpoint or lighting that were less relevant. To address this issue, a dual attention mechanism-based image change captioning method was introduced in 2019, aimed at accurately identifying the areas of change between images[26)]. However, both methods assume that differences between the two inputs will manifest in localized regions of the signal (in the case of images, in rectangular regions). Unfortunately, this assumption does not necessarily hold true for the sounds of machine operation, which include components that span wide ranges of the time and frequency domains. Moreover, the existing techniques for image change captioning are not easily applied to acoustic change captioning. Instead, Hitachi developed a new acoustic change captioning technique that was informed by the image change captioning method based on the dual attention mechanism[26)].

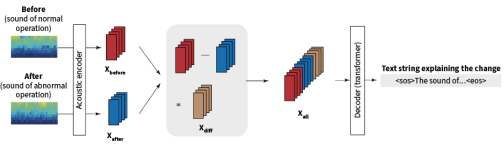
3. Proposed Method

3.1 AI Model for Acoustic Change Captioning

Hitachi has proposed an AI model for acoustic change captioning that can automatically generate text explaining the changes between two (before and after) acoustic signal inputs. The model has an encoder-decoder neural network architecture made up of an acoustic encoder that treats these before and after acoustic signals as variable-length vectors and encodes them as a single vector with fixed dimensionality and a decoder that decodes this encoded vector as a text string.

The model starts by obtaining the logarithmic mel scale spectrograms for each of the input signals by taking the Fast Fourier Transform (FFT) of each signal, calculating the power spectrograms, and converting to a logarithmic mel scale in the frequency axis. These spectrograms are then input into the same acoustic encoder to generate two fixed-dimension vectors that encode the before and after acoustic signals (X_{before} and X_{after}). Working in the same way as the image change captioning technique proposed in 2019, the AI model calculates the difference between the two vectors (X_{diff}) to make this difference easier to work with. The three vectors X_{before}, X_{after}, and X_{diff} are then concatenated into a combined vector X_{all} that is input to the decoder to generate the text explaining the change (see Figure 1).

Figure 1—AI Model for Acoustic Change Captioning



The logarithmic mel scale spectrograms for the before and after acoustic signals are input into the acoustic encoder to obtain a combined vector made up of the vectors for each of the signals and a difference vector. This combined vector is then input to the decoder to generate the text string explaining the change.

Two different acoustic encoders are used, a standard transformer encoder[27)] and a spatial attention encoder like that used in the image change captioning technique based on the dual attention mechanism[26)]. The utility of conventional transformer encoders for acoustic captioning has already been demonstrated. The spatial attention technique, on the other hand, uses a convolutional neural network and directs attention at localized regions of the time and frequency domains on the assumption that the changes from one acoustic signal to the next will be localized in this way. While this assumption does not always hold true for the sounds of machine operation, as noted above, it is anticipated that the spatial attention technique will function well for stationary sound, which in many cases does exhibit such localized behavior in the frequency direction. In this way, separate architectures are used on the basis that the optimal architecture will likely differ depending on which aspect of machine operation sound is to be explained. That is, three such aspects were defined (stationary sound, periodic sound, and non-periodic sound), with spatial attention being used as the acoustic encoder for stationary sound and the transformer encoder for periodic and non-periodic sound (both being non-stationary sounds).

For the decoder, meanwhile, a standard transformer model is used, a technique that has already been demonstrated as effective for general acoustic captioning. This model is made up of a multi-head encoder-decoder attention layer and a multi-head self-attention layer for the text string.

3.2 Data Set and Model Training

As there were no existing data sets for training and testing acoustic change captioning, Hitachi created its own. Named MIMII-Change, the data set was based on the Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization (MIMII-DG) data set[17] developed for training and testing abnormal sound detection. MIMII-Change provides operating sounds from five different types of machines (bearings, fans, gearboxes, sliders, and valves). It contains 300 acoustic signal pairs, each of which contains the sound of a particular machine operating normally and abnormally. Each pair is also accompanied by text explaining the change in sound from normal to abnormal operation.

These explanatory texts were collated manually by three annotators. The annotators were instructed to use onomatopoeic words wherever possible to describe the changes in sound. An onomatopoeia is a word or words that phonetically imitate the sound being described. They provide a useful way of characterizing the diverse sounds that occur in the environment[28]. Moreover, by having the annotators assign separate explanatory text to stationary, periodic, and non-periodic sound respectively, the different model architectures can each be trained using the relevant explanatory text.

As with generic acoustic captioning, model training can be performed on the basis of cross-entropy loss. That is, the acoustic encoder and decoder are trained to make the output text for each pair match the corresponding explanatory text as closely as possible. Here, the output text means the text obtained when the before and after acoustic signal pairs from the training data are input to the system.

4. Performance Testing

Table 1 lists the results of testing on stationary, periodic, and non-periodic sounds. BLEU, METEOR, CIDEr, SPICE, and SPIDEr are commonly used metrics for captioning, with higher values indicating closer agreement between the output text and the correct explanation text. MPER, meanwhile, is a recently defined metric for assessing the accuracy of the onomatopoeic parts of text. It is the mean value of the phoneme error rate for the combined imitation sounds present in a text, with lower values indicating that the phoneme strings in the onomatopoeic parts of the output text more closely resemble the correct explanation text.

Table 1 lists the testing results. These indicate that the acoustic encoder based on spatial attention is suitable for explaining changes in stationary sound. Similarly, the results also show that the transformer encoder is suitable for explaining changes in non-stationary (periodic and non-periodic) sound. This indicates that it is desirable to switch between these two encoder models depending on whether it is the change in stationary or non-stationary sound that is being explained.

Table 1—Testing Results

Aspect of machine sound being explained	Encoder	BLEU 3	BLEU 4	METEOR	CIDEr	SPICE	SPIDEr	MPER
Stationary sound	Transformer	0.616	0.542	0.427	0.969	0.34	0.655	0.281
	Spatial attention	**0.669**	**0.601**	**0.441**	**1.086**	**0.365**	**0.726**	**0.266**
Periodic sound	Transformer	**0.464**	**0.387**	0.39	**0.946**	**0.255**	**0.601**	**0.338**
	Spatial attention	0.426	0.354	**0.402**	0.881	0.249	0.565	0.38
Non-periodic sound	Transformer	**0.413**	**0.339**	**0.427**	**1.864**	**0.373**	**1.118**	0.327
	Spatial attention	0.328	0.269	0.411	1.441	0.304	0.873	**0.321**

The table lists the results for a range of metrics of using the two encoders for stationary, periodic, and non-periodic sound. A higher value indicates closer agreement with the correct explanatory text for all metrics except MPER, where the opposite applies. Figures highlighted in bold indicate which of the encoders gave a better result.

5. Conclusions

This article has described an acoustic change captioning technique that is used after an AI has detected an abnormal sound. The technique generates text explaining how the character of the sound changed between the normal and abnormal sounds. Hitachi plans to further develop this technique for use in automating acoustic inspection and expanding its scope of application.

The technique is one of a package of AI techniques for interpreting what is happening in a workplace and providing information to the on-site staff that will help them decide what action to take. It is anticipated that the development of these techniques will lead to the implementation of AI agents that can provide interactive support for the work of on-site staff.

Acknowledgements

In the development of the acoustic change captioning described in this article, Hitachi received assistance from Shunsuke Tsubaki (at that time a master's student) and Professor Keisuke Imoto of Doshisha University, and Yuki Okamoto (at that time a doctoral student) of Ritsumeikan University. The authors would like to take this opportunity to express their deep gratitude.

REFERENCES

1) Hitachi, Ltd. "Retrofitted Wireless Sensors (Automatic Meter Reading and Anomaly Detection)," in Japanese.
2) Hitachi, Ltd. "Hitachi Intelligent Platform Use Cases, Maintenance DX," in Japanese.
3) H. Purohit et al., "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2019)
4) R. Tanabe et al., "MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2021)
5) K. Dohi et al., "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2022)
6) Y. Koizumi et al., "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)
7) Y. Kawaguchi et al., "Description and discussion on DCASE 2021 Challenge Task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2021)
8) 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. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2022)
9) K. Dohi et al., "Description and discussion on DCASE 2023 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2023)
10) T. Nishida et al., "Description and discussion on DCASE 2024 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2024)
11) Y. Kawaguchi et al., "How can we detect anomalies from subsampled audio signals?," in Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (2017)
12) Y. Kawaguchi, "Anomaly detection based on feature reconstruction from subsampled audio signals," in Proc. European Signal Processing Conference (EUSIPCO) (2018)
13) Y. Kawaguchi et al., "Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019)
14) K. Suefusa et al., "Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)
15) H. Purohit et al., "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)
16) K. Dohi et al., "Flow-based self-supervised density estimation for anomalous sound detection," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021)
17) K. Dohi et al., "Disentangling physical parameters for anomalous sound detection under domain shifts," in Proc. European Signal Processing Conference (EUSIPCO) (2022)
18) T. Nishida et al., "Anomalous sound detection based on machine activity detection," in Proc. European Signal Processing Conference (EUSIPCO) (2022)
19) H. Purohit et al., "Hierarchical conditional variational autoencoder based acoustic anomaly detection," in Proc. European Signal Processing Conference (EUSIPCO) (2022)
20) K. Shimonishi et al., "Anomalous sound Detection based on sound separation," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH) (2023)
21) K. Dohi et al., "Distributed collaborative anomalous sound detection by embedding sharing," in Proc. European Signal Processing Conference (EUSIPCO) (2024)
22) T.V. Ho et al., "Stream-based active learning for anomalous sound detection in machine condition monitoring," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH) (2024)
23) K. Imoto et al., "Research Trends in Environmental Sound Analysis and Anomalous Sound Detection," IEICE ESS Fundamentals Review, Engineering Sciences Society, The Institute of Electronics, Information and Communication Engineers (2022)
24) S. Tsubaki et al., "Audio-change captioning to explain machine-sound anomalies," in Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2023)

25) H. Jhamtani et al., “Learning to describe differences between pairs of similar images,” in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)

26) D.H. Park et al., “Robust change captioning,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

27) A. Vaswani et al., “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NIPS) (2017)

28) Y. Okamoto et al., “Environmental sound extraction using onomatopoeic words,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022)

© Hitachi, Ltd. 1994, 2025. All rights reserved.

Hitachi Review

Hitachi Review is a technical medium that reports on Hitachi's use of innovation to address the challenges facing society.

The *Hitachi Review* website contains technical papers written by Hitachi engineers and researchers, special articles such as discussions or interviews, and back numbers.

Hitachi Hyoron (Japanese) website	*Hitachi Review* (English) website
https://www.hitachihyoron.com/jp/	https://www.hitachihyoron.com/rev/

Hitachi Review Newsletter

Hitachi Review newsletter delivers the latest information about Hitachi Review when new articles are released.