IEICE TRANSACTIONS

on Information and Systems

VOL. E104-D NO. 3
MARCH 2021

The usage of this PDF file must comply with the IEICE Provisions on Copyright.
The author(s) can distribute this PDF file for research and educational (nonprofit) purposes only.
Distribution by anyone other than the author(s) is prohibited.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Noise Robust Acoustic Anomaly Detection System with Nonnegative Matrix Factorization Based on Generalized Gaussian Distribution*

Akihito AIBA[†a)], **Minoru YOSHIDA**[†], ***Nonmembers***, **Daichi KITAMURA**[††,†††], ***Member***, **Shinnosuke TAKAMICHI**[†††], ***Nonmember***, ***and*** **Hiroshi SARUWATARI**[†††], ***Member***

SUMMARY We studied an acoustic anomaly detection system for equipments, where the outlier detection method based on recorded sounds is used. In a real environment, the SNR of the target sound against background noise is low, and there is the problem that it is necessary to catch slight changes in sound buried in noise. In this paper, we propose a system in which a sound source extraction process is provided at the preliminary stage of the outlier detection process. In the proposed system, nonnegative matrix factorization based on generalized Gaussian distribution (GGD-NMF) is used as a sound source extraction process. We evaluated the improvement of the anomaly detection performance in a low-SNR environment. In this experiment, SNR capable of detecting an anomaly was greatly improved by providing GGD-NMF for preprocessing.

key words: *nonnegative matrix factorization, generalized Gaussian distribution, anomaly detection, outlier detection*

1. Introduction

Several systems have been proposed with which abnormal operation is detected from sensor signals by constantly sensing the state of facilities such as water supply and drainage and air conditioning in place of human beings [1], [2]. Here, we consider an anomaly detection system based on the operating sound recorded by a microphone.

One of the major problems in anomaly detection applications is the difficulty in obtaining samples at anomalous times. In particular, it is difficult to intentionally create an anomalous state of a facility that is expensive and important for building management, for the purpose of sample collection. Also, waiting for a naturally occurring anomaly is impractical as some anomalies occur only once every few months or years. For this reason, outlier detection is often adopted in anomaly detection applications [3], [4]. In outlier detection, only easily available normal samples are collected in advance. Then their features are modeled, and samples that deviate from the model are judged to be anomalies when obtained. In this research, we also use this outlier detection approach.

Manuscript received May 18, 2020.
Manuscript revised October 14, 2020.
Manuscript publicized December 18, 2020.
[†]The authors are with Ricoh Company Ltd., Ebina-shi, 243–0460 Japan.
[††]The author is with the National Institute of Technology, Kagawa College, Takamatsu-shi, 761–8058 Japan.
[†††]The authors are with the University of Tokyo, Tokyo, 113–8656 Japan.
*This is a paper on system development.
a) E-mail: akihito.aiba@jp.ricoh.com
DOI: 10.1587/transinf.2020EDK0002

When sound data is used as input, the background noise in a real environment also becomes a problem. We often see a situation where multiple equipment units are installed in one room and generate very loud operation noise. In such a situation, even if the operation sound of one device shows an abnormal feature, its detection is very difficult because the signal power of the anomalous sound is lower than that of the operation sound of other devices. This is a common problem not only in outlier detection but also in sound discrimination. There are existing methods for improving the SNR by extracting sound sources before discrimination [5], [6].

In this paper, a method based on nonnegative matrix factorization (NMF) [7] is used for sound source extraction processing. Regarding the sound source extraction, deep neural network (DNN)-based processing [8]–[10] has been proposed for general use in recent years. However, it cannot be introduced in our anomaly detection application because DNN requires an enormous number of clean objective sounds but we cannot obtain them in advance. On the other hand, the proposed NMF-based method has an advantage to requiring no clean data. In particular, we extend the NMF based on generalized Gaussian distribution (GGD-NMF) [11] proposed by Kitamura to a semisupervised NMF (SS-NMF) [12], [13] and confirm its efficacy in the anomaly detection system. GGD-NMF incorporates a generative model based on GGD into NMF to control the low-rank property during decomposition. The evaluation is performed using data recorded in a real environment.

This paper is organized as follows. In Sect. 2, we introduce the basis theory of NMF and its application to audio signals. In Sect. 3, the details of the proposed anomaly detection system are explained. The experimental evaluation of the proposed system is described in Sect. 4. The paper is concluded in Sect. 5.

2. Conventional NMF

In this section, the NMF algorithm is described first. Next, Itakura–Saito NMF (ISNMF) [14], which has a Gaussian-distribution-based generative model, is introduced.

2.1 NMF

NMF is a mathematical algorithm for extracting a limited number of nonnegative features from a nonnegative matrix.

Copyright © 2021 The Institute of Electronics, Information and Communication Engineers

Let $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{I \times J}$ be a nonnegative observation matrix to be analyzed. In the field of acoustics, the complex spectrogram $\boldsymbol{C} \in \mathbb{C}^{I \times J}$ is obtained by applying short-time Fourier transform (STFT) to the time-domain signal. Then, the nonnegative spectrogram $\boldsymbol{X}$ is calculated from $\boldsymbol{C}$, e.g., the amplitude spectrogram ($\boldsymbol{X} = |\boldsymbol{C}|^{.1}$) or the power spectrogram ($\boldsymbol{X} = |\boldsymbol{C}|^{.2}$), where $|\cdot|^{.p}$ returns a matrix with the element-wise absolute value and the pth power of the input matrix. I and J are the numbers of rows and columns in $\boldsymbol{X}$, respectively. In the case of a spectrogram, I and J correspond to the numbers of frequency bins and time frames, respectively.

NMF is an approximate decomposition as follows:

$$\boldsymbol{X} \approx \hat{\boldsymbol{X}} = \boldsymbol{F}\boldsymbol{G} \tag{1}$$

$$= \sum_k \boldsymbol{f}_k \boldsymbol{g}_k^T, \tag{2}$$

where $\boldsymbol{F} = (\boldsymbol{f}_1 \ldots \boldsymbol{f}_K) \in \mathbb{R}_{\geq 0}^{I \times K}$ is the basis matrix, and $\boldsymbol{G} = (\boldsymbol{g}_1 \ldots \boldsymbol{g}_K)^T \in \mathbb{R}_{\geq 0}^{K \times J}$ is the coefficient matrix or activation matrix. $k = 1, 2, \ldots, K$ is the index of bases. K is the number of bases and is generally set to a value sufficiently smaller than I and J. Therefore, Eq. (1) is interpreted as a low-rank approximation that represents the observation matrix $\boldsymbol{X}$ with a limited number of bases. The basis vectors $\boldsymbol{f}_k \in \mathbb{R}_{\geq 0}^{I}$ represent the frequently appearing spectral patterns (parts) in the spectrogram $\boldsymbol{X}$, and the activation vectors $\boldsymbol{g}_k \in \mathbb{R}_{\geq 0}^{J}$ represent their time-varying gains [11].

The matrix variables $\boldsymbol{F}$ and $\boldsymbol{G}$ can be estimated by solving the following optimization problem:

$$\min_{\boldsymbol{F},\boldsymbol{G}} \mathcal{D}(\boldsymbol{X}|\boldsymbol{F}\boldsymbol{G}) \text{ s.t. } f_{ik}, g_{kj} \geq 0 \ \forall i, j, k, \tag{3}$$

where $i = 1, 2, \ldots, I$ is the index of frequency bins and $j = 1, 2, \ldots, J$ is the index of time frames. f_{ik} and g_{kj} are the elements of $\boldsymbol{F}$ and $\boldsymbol{G}$, respectively. $\mathcal{D}(\cdot|\cdot)$ is a similarity function between two input matrices. For example, squared Euclidean distance [15], generalized Kullbak–Leibler (KL) divergence [15], and Itakura–Saito (IS) divergence [14] are often used.

2.2 ISNMF

In this section, we consider NMF based on IS divergence for the similarity function $\mathcal{D}$ in Eq. (3) (ISNMF). IS divergence is a similarity function defined as

$$\mathcal{D}_{\mathrm{IS}}(c|\sigma) = \frac{|c|^2}{\sigma^2} - \log \frac{|c|^2}{\sigma^2} - 1. \tag{4}$$

By setting $|c|^2 = x_{ij}$ and $\sigma^2 = \sum_k f_{ik} g_{kj}$, we can rewrite the optimization problem of ISNMF as

$$\min_{\boldsymbol{F},\boldsymbol{G}} \sum_{i,j} \left(\frac{x_{ij}}{\sum_k f_{ik} g_{kj}} + \log \sum_k f_{ik} g_{kj} \right) \text{ s.t. } f_{ik}, g_{kj} \geq 0 \ \forall i, j, k, \tag{5}$$

where x_{ij} is the element of the complex spectrogram $\boldsymbol{X}$ and the constant terms are omitted.

In ISNMF, a generative model of the complex spectrogram $\boldsymbol{C}$ is assumed as described below. Let us assume that each time-frequency element of the complex spectrogram $\boldsymbol{C}$, c_{ij}, can be decomposed into K spectral components $c_{ij,k}$ as $c_{ij} = \sum_k c_{ij,k}$, and $c_{ij,k}$ is assumed to be generated from a zero-mean circularly symmetric complex Gaussian distribution with the variance $\sigma_{ij,k}^2 > 0$ [14]. Then, the generative model of the complex spectrogram $\boldsymbol{C}$ becomes

$$\boldsymbol{C} \sim \prod_{i,j} p(c_{ij}) = \prod_{i,j} \mathcal{N}_{\mathbb{C}} \left(0, \sum_k \sigma_{ij,k}^2 \right). \tag{6}$$

On the basis of this generative model, we consider the maximum likelihood estimation problem of the variance $\sigma_{ij,k}^2$. The negative log-likelihood function of $\boldsymbol{C}$ is

$$-\log \mathcal{L} = -\log \prod_{i,j} \mathcal{N}_{\mathbb{C}} \left(0, \sum_k \sigma_{ij,k}^2 \right) = \sum_{i,j} \left(\frac{|c_{ij}|^2}{\sum_k \sigma_{ij,k}^2} + \log \sum_k \sigma_{ij,k}^2 + \log \pi \right). \tag{7}$$

The maximum likelihood estimate of the variance $\sigma_{ij,k}^2$ can be obtained by minimizing Eq. (7). Here, when comparing Eq. (7) with the ISNMF minimization problem (5), if $x_{ij} = |c_{ij}|^2$ and $\sigma_{ij,k}^2 = f_{ik} g_{kj}$, both equations become equivalent up to the constant term.

From the above, it is seen that applying ISNMF to the power spectrogram $\boldsymbol{X} = |\boldsymbol{C}|^{.2}$ is equivalent to the maximum likelihood estimation of the variance $\sigma_{ij,k}^2$ based on the generative model (6).

3. Proposed System

In this section, GGD-NMF is described first. Then, we describe the proposed system that extends GGD-NMF to SS-NMF and incorporates it into the anomaly detection system.

3.1 GGD-NMF

In ISNMF, a complex Gaussian distribution is assumed as a generative model of the observation complex spectrogram $\boldsymbol{C}$. GGD-NMF is a generalization of ISNMF, which assumes the complex GGD as a generative model of $\boldsymbol{C}$.

By denoting the circularly symmetric complex GGD with the mean value 0, shape parameter $\rho > 0$, and scale parameters σ_{ij} that fluctuate along time and frequency as $\mathcal{G}_{\mathbb{C}}(c_{ij}; 0, \rho, \sigma_{ij})$, we can describe the generative model in GGD-NMF as

$$\boldsymbol{C} \sim \prod_{i,j} \mathcal{G}_{\mathbb{C}}(c_{ij}; 0, \rho, \sigma_{ij}) = \prod_{i,j} \frac{\rho^{1-\frac{2}{\rho}}}{2^{1-\frac{2}{\rho}} \pi \sigma_{ij}^2 \Gamma(2/\rho)} \exp\left[-\frac{2}{\rho} \left(\frac{|c_{ij}|}{\sigma_{ij}} \right)^{\rho} \right], \tag{8}$$

$$\sigma_{ij}^{p} = \sum_k f_{ik} g_{kj}, \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function and p is a domain parameter corresponding to the domain of the low-rank approximation. In the case of $p = 1$, the amplitude spectrogram $|\boldsymbol{C}|^{\cdot 1}$ is modeled by NMF variables ($\boldsymbol{F}$ and $\boldsymbol{G}$), and in the case of $p = 2$, the power spectrogram $|\boldsymbol{C}|^{\cdot 2}$ is modeled. The complex GGD $\mathcal{G}_{\mathbb{C}}(c_{ij}; 0, \rho, \sigma_{ij})$ becomes equivalent to the complex Gaussian distribution and the complex Laplace distribution when $\rho = 2$ and $\rho = 1$, respectively. Also, the complex GGD becomes sub- and super-Gaussian when $\rho > 2$ and $\rho < 2$, respectively.

In the generative model (8), the complex GGD is independently defined in each time-frequency slot, and its scale parameter σ_{ij} can fluctuate along time and frequency axes. In this case, the macro model that includes all the time-frequency slots, i.e., the generative model of the spectrogram $\boldsymbol{C}$, always approaches to a super-Gaussian distribution when σ_{ij} dynamically varies along i and j [16]. For this reason, the generative model (8) with sub-Gaussian GGD ($\rho > 2$) has a versatility as depicted in Fig. 1 in [16], namely, the macro model of Eq. (8) with $\rho > 2$ becomes either sub-Gaussian or super-Gaussian. This versatile model is desirable and often provides improved performance in NMF-based acoustic modeling [16], [17].

Let us clarify the relationship between the generative model (8) and the similarity function in NMF. The divergence based on the complex GGD is derived from the difference in log-likelihood (deviance). The divergence can be calculated as follows [11]:

$$\begin{aligned}\mathcal{D}_{\mathrm{GGD}}(c\|\sigma) &= -2\log|c| - \frac{2}{\rho} + 2\log\sigma + \frac{2}{\rho}\left(\frac{|c|}{\sigma}\right)^{\rho} \\ &= \frac{2}{\rho}\left[\left(\frac{|c|}{\sigma}\right)^{\rho} - \log\left(\frac{|c|}{\sigma}\right)^{\rho} - 1\right]. \end{aligned} \tag{10}$$

This divergence (10) can be interpreted as a generalization of IS divergence (4) for the shape parameter ρ. From the above, the function to be minimized in GGD-NMF is as follows (except for the constant term):

$$\sum_{i,j}\mathcal{D}_{\mathrm{GGD}}(c_{ij}|\sigma_{ij}) = \sum_{i,j}\left[\frac{|c_{ij}|^{\rho}}{\left(\sum_k f_{ik}g_{kj}\right)^{\frac{\rho}{p}}} + \frac{\rho}{p}\log\sum_k f_{ik}g_{kj}\right]. \tag{11}$$

Thus, GGD-NMF (Eq. (11)) and ISNMF (Eq. (5)) are equivalent algorithms when $\rho = p$ and $x_{ij} = |c|^p$.

The variables f_{ik} and g_{kj} that minimize Eq. (11) can be calculated by the auxiliary function technique. The update rules are derived as follows [11], [18]:

$$f_{ik} \leftarrow f_{ik}\left[\frac{\sum_j \frac{z_{ij}}{\left(\sum_{k'} f_{ik'}g_{k'j}\right)^2}g_{kj}}{\sum_j \frac{1}{\sum_{k'} f_{ik'}g_{k'j}}g_{kj}}\right]^{\frac{p}{\rho+p}}, \tag{12}$$

$$g_{kj} \leftarrow g_{kj}\left[\frac{\sum_i \frac{z_{ij}}{\left(\sum_{k'} f_{ik'}g_{k'j}\right)^2}f_{ik}}{\sum_i \frac{1}{\sum_{k'} f_{ik'}g_{k'j}}f_{ik}}\right]^{\frac{p}{\rho+p}}, \tag{13}$$

$$z_{ij} = \left(|c_{ij}|^{\frac{\rho}{p}}\,\sigma_{ij}^{1-\frac{\rho}{p}}\right)^{p}. \tag{14}$$

In the update rules, z_{ij} is regarded as a virtual observation consisting of the weighted geometric mean of the observation $|c_{ij}|$ and its low-rank model σ_{ij}, and GGD-NMF is interpreted as ISNMF on this virtual observation. When $\rho = p$, the virtual observation becomes $z_{ij} = |c_{ij}|^p$, which coincides with the original ISNMF. When $\rho < p$, since the virtual observation z_{ij} is the weighted geometric mean of $|c_{ij}|$ and σ_{ij}, GGD-NMF tends to prevent σ_{ij} from an overfitting to the observation $|c_{ij}|$. This is because a heavy-tailed (super-Gaussian) distribution is assumed and it allows outliers in $|c_{ij}|$, resulting in low-rank-enhanced NMF modeling. On the other hand, when $\rho > p$, the geometric mean corresponds to the point externally dividing $|c_{ij}|$ and σ_{ij}, and the error with the current approximation is further emphasized, which mitigates the excessive low-rank modeling [11]. Thus, in GGD-NMF, it is possible to control the low-rank property of the approximated model by adjusting the parameters ρ and p.

3.2 GGD-Based SS-NMF

In SS-NMF [12], [13], we have a sample sound (training dataset) of one source, and its basis matrix is trained in advance by decomposing the sample sound using simple NMF. Then, the observed mixture spectrogram $\boldsymbol{X}$ is divided into the components with the pretrained (supervised) basis matrix and the other basis matrix as

$$\begin{aligned}\boldsymbol{X} &\approx \boldsymbol{FG} + \boldsymbol{HU} \\ &= \sum_k \boldsymbol{f}_k\boldsymbol{g}_k + \sum_l \boldsymbol{h}_l\boldsymbol{u}_l, \end{aligned} \tag{15}$$

where $\boldsymbol{F}$ is the pretrained basis matrix that has spectral patterns of the target source, and $\boldsymbol{H} = (\boldsymbol{h}_1 \ldots \boldsymbol{h}_L) \in \mathbb{R}_{\geq 0}^{I\times L}$ is the other basis matrix estimated in the separation stage and represents the spectral patterns of the other sources in $\boldsymbol{X}$. $\boldsymbol{G}$ and $\boldsymbol{U} = (\boldsymbol{u}_1 \ldots \boldsymbol{u}_L)^T \in \mathbb{R}_{\geq 0}^{L\times J}$ are the coefficient matrices for $\boldsymbol{F}$ and $\boldsymbol{H}$, respectively, and both are also estimated in the separation stage. L is the number of other bases in $\boldsymbol{H}$, and $l = 1, 2, \ldots, L$ is their index.

The matrices $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ can be estimated by the following optimization problem, which is the separation stage of SS-NMF.

$$\begin{aligned}\min_{\boldsymbol{H},\boldsymbol{G},\boldsymbol{U}} \quad & \mathcal{D}(\boldsymbol{X}|\boldsymbol{FG} + \boldsymbol{HU}) \\ \text{s.t. } & g_{kj}, h_{il}, u_{lj} \geq 0 \ \forall i, j, k, l \end{aligned} \tag{16}$$

Since $\boldsymbol{F}$ contains the spectral patterns of the target source and is fixed in the optimization of Eq. (16), $\boldsymbol{FG}$ and $\boldsymbol{HU}$ correspond to the estimated nonnegative spectrograms of the target source and the other source components, respectively. Note that the similarity function $\mathcal{D}(\cdot|\cdot)$ must be the same as that used in the training of $\boldsymbol{F}$.

In the proposed system, we introduce GGD-NMF into the framework of SS-NMF. The function to be minimized changes from Eq. (11) as follows:

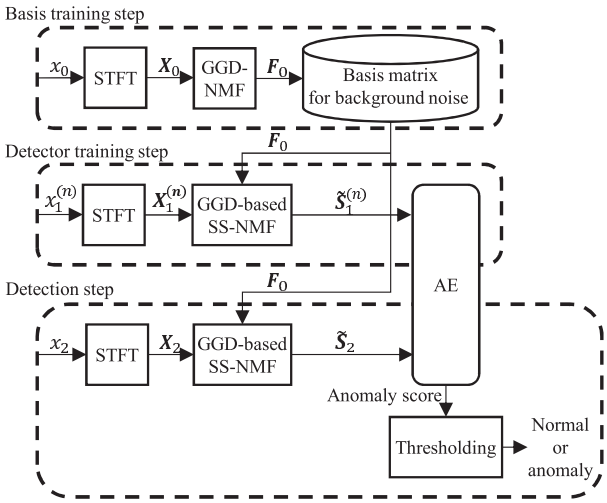


Fig. 1 System overview.

$$\sum_{i,j} \mathcal{D}_{\mathrm{GGD}}(c_{ij}|\sigma_{ij}) = \sum_{i,j}\left[\left(\frac{|c_{ij}|}{\sigma_{ij}}\right)^{\rho} + \rho \log \sigma_{ij}\right], \tag{17}$$

$$\sigma_{ij}^{p} = \sum_{k} f_{ik} g_{kj} + \sum_{l} h_{il} u_{lj}. \tag{18}$$

The variables that minimize Eq. (17) can be calculated by the auxiliary function technique similarly to the derivation of Eqs. (12) and (13) [18]. The update rules for solving GGD-based SS-NMF are derived as

$$g_{kj} \leftarrow g_{kj}\left[\frac{\sum_i \frac{z_{ij}}{\left(\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}\right)^2} f_{ik}}{\sum_i \frac{1}{\left(\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}\right)} f_{ik}}\right]^{\frac{p}{\rho+p}}, \tag{19}$$

$$h_{il} \leftarrow h_{il}\left[\frac{\sum_j \frac{z_{ij}}{\left(\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}\right)^2} u_{lj}}{\sum_j \frac{1}{\left(\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}\right)} u_{lj}}\right]^{\frac{p}{\rho+p}}, \tag{20}$$

$$u_{lj} \leftarrow u_{lj}\left[\frac{\sum_i \frac{z_{ij}}{\left(\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}\right)^2} h_{il}}{\sum_i \frac{1}{\left(\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}\right)} h_{il}}\right]^{\frac{p}{\rho+p}}, \tag{21}$$

$$z_{ij} = \left(|c_{ij}|^{\frac{\rho}{p}} \sigma_{ij}^{1-\frac{\rho}{p}}\right)^{p}. \tag{22}$$

3.3 System Overview

Figure 1 shows the overview of the proposed system using GGD-NMF for anomaly detection. GGD-NMF is expected to be effective in improving the anomaly detection performance even in a low-SNR environment. The proposed system consists of the following three major steps.

1. Basis training step: training the basis matrix $\boldsymbol{F}_0$ for background noise.
2. Detector training step: training to learn the normal operation sound of a target machine, which is extracted by GGD-based SS-NMF.
3. Detection step: detecting anomalies from the extracted machine-operation sound.

The details of each step are described below.

3.4 Basis Training Step

In the proposed system, first, we obtain the basis matrix of background noise. A time-domain sound signal x_0 is recorded when the target machine is not operating and only background noise exists. The spectrogram of x_0, $\boldsymbol{X}_0 \in \mathbb{R}_{\geq 0}^{\Omega \times T_0}$ is obtained by STFT, where Ω and T_0 are the numbers of frequency bins and time frames, respectively. It can be considered as a noise sample spectrogram $\boldsymbol{N}_0 \in \mathbb{R}_{\geq 0}^{\Omega \times T_0}$ because $\boldsymbol{X}_0$ does not include the operation sound of the target machine. Then, $\boldsymbol{N}_0$ is decomposed by GGD-NMF as

$$\boldsymbol{X}_0 = \boldsymbol{N}_0 \approx \boldsymbol{F}_0 \boldsymbol{W}_0, \tag{23}$$

where $\boldsymbol{F}_0 \in \mathbb{R}_{\geq 0}^{\Omega \times K}$ is a basis matrix for background noise and $\boldsymbol{W}_0 \in \mathbb{R}_{\geq 0}^{K \times T_0}$ is a coefficient matrix for $\boldsymbol{F}_0$. As a result, the basis matrix that represents only background noise sepectra is obtained in this training process.

3.5 Detector Training Step

In the detector training step, the operation sound of the target machine is extracted from the noisy mixture sound, and the extracted sound is used to train the anomaly detector. To extract the operation sound of the target machine, the basis matrix $\boldsymbol{F}_0$ obtained in the basis training step is used.

We prepare a noisy sound dataset $\{x_1^{(1)}, x_1^{(2)}, \ldots, x_1^{(N)}\}$ recorded when the target machine was operating normally in a noisy environment. Their nonnegative spectrograms $\{\boldsymbol{X}_1^{(1)}, \boldsymbol{X}_1^{(2)}, \ldots, \boldsymbol{X}_1^{(N)}\}$ are calculated by STFT, where $\boldsymbol{X}_1^{(n)} \in \mathbb{R}_{\geq 0}^{\Omega \times T_1^{(n)}}$, $T_1^{(n)}$ is the number of time frames, N is the number of samples in the dataset, and $n = 1, 2, \ldots, N$ is their index. We assume that $\boldsymbol{X}_1^{(n)}$ can be approximated as the sum of two nonnegative spectrograms as

$$\boldsymbol{X}_1^{(n)} \simeq \boldsymbol{S}_1^{(n)} + \boldsymbol{N}_1^{(n)}, \tag{24}$$

where $\boldsymbol{S}_1^{(n)} \in \mathbb{R}_{\geq 0}^{\Omega \times T_1^{(n)}}$ and $\boldsymbol{N}_1^{(n)} \in \mathbb{R}_{\geq 0}^{\Omega \times T_1^{(n)}}$ are the nonnegative spectrograms of the clean operation sound of the target machine and the background noise, respectively.

In the detector training step, first, the observed noisy mixture $\boldsymbol{X}_1^{(n)}$ is decomposed into two estimates, i.e., the background noise $\tilde{\boldsymbol{N}}_1^{(n)}$ and the operation sound of the target machine $\tilde{\boldsymbol{S}}_1^{(n)}$, by GGD-based SS-NMF with the pretrained basis matrix $\boldsymbol{F}_0$ as follows:

$$\boldsymbol{X}_1^{(n)} \simeq \tilde{\boldsymbol{S}}_1^{(n)} + \tilde{\boldsymbol{N}}_1^{(n)}, \tag{25}$$

$$\tilde{\boldsymbol{S}}_1^{(n)} = \boldsymbol{H}_1^{(n)} \boldsymbol{U}_1^{(n)}, \tag{26}$$

$$\tilde{\boldsymbol{N}}_1^{(n)} = \boldsymbol{F}_0 \boldsymbol{G}_1^{(n)}, \tag{27}$$

where $\boldsymbol{H}_1^{(n)} \in \mathbb{R}_{\geq 0}^{\Omega \times L}$ and $\boldsymbol{U}_1^{(n)} \in \mathbb{R}_{\geq 0}^{L \times T_1^{(n)}}$ are the basis and coefficient matrices that correspond to the operation sound of the target machine as Eq. (26) and $\boldsymbol{G}_1^{(n)} \in \mathbb{R}_{\geq 0}^{K \times T_1^{(n)}}$ is the coefficient matrix of $\boldsymbol{F}_0$, which corresponds to the background

noise as Eq. (27).

Next, the anomaly detector is trained using the extracted normal operation sound $\tilde{\boldsymbol{S}}_1^{(n)}$ as the training data. In the proposed system, we use an outlier detection method with Autoencoder (AE) [4]. AE is a neural network trained to reconstruct input data through dimensional compression. For this reason, the reconstruction error of the trained AE becomes smaller for known inputs, whereas the reconstruction error becomes larger for unknown inputs. Thus, it is possible to detect outliers on the basis of the reconstruction error.

A parameter of AE, Θ, is optimized by minimizing a loss function $\mathcal{J}$:

$$\mathcal{J}(\Theta) = \mathbb{E}\left[\left\|\boldsymbol{S}_1'^{(n,m)} - f_{\mathrm{AE}}\left\{\boldsymbol{S}_1'^{(n,m)}\middle|\Theta\right\}\right\|_2^2\right]_{n,m}, \tag{28}$$

$$\boldsymbol{S}_1'^{(n,m)} = \left|\tilde{\boldsymbol{S}}_1^{(n,m)}\right|^{\cdot\frac{p_{\mathrm{AE}}}{p}}, \tag{29}$$

where $\tilde{\boldsymbol{S}}_1^{(n,m)} \in \mathbb{R}_{\geq 0}^{\Omega\times T_{\mathrm{AE}}}$ is a spectrogram piece cut out from $\tilde{\boldsymbol{S}}_1^{(n)}$ every time frame number T_{AE}, $m = 1, 2, \ldots, M_n$ is the index of cut spectrogram pieces, and M_n is the number of cut spectrogram pieces for the nth sample sound. Function $f_{\mathrm{AE}}\{\cdot \mid \Theta\}$ represents the processing of AE with the given parameter Θ and p_{AE} is a domain parameter used in AE.

3.6 Detection Step

In the detection step, a sound signal x_2 is recorded when the target machine is operating in a noisy environment. Its nonnegative spectrogram $\boldsymbol{X}_2$ is obtained by STFT, where $\boldsymbol{X}_2 \in \mathbb{R}_{\geq 0}^{\Omega\times T_2}$, and T_2 is the number of time frames. The operation sound of the target machine is extracted from $\boldsymbol{X}_2$ by GGD-based SS-NMF with $\boldsymbol{F}_0$, as follows:

$$\boldsymbol{X}_2 \simeq \boldsymbol{F}_0\boldsymbol{G}_2 + \boldsymbol{H}_2\boldsymbol{U}_2, \tag{30}$$

$$\tilde{\boldsymbol{S}}_2 = \boldsymbol{H}_2\boldsymbol{U}_2, \tag{31}$$

where $\tilde{\boldsymbol{S}}_2 \in \mathbb{R}_{\geq 0}^{\Omega\times T_2}$ is the nonnegative spectrogram of a clean operation sound of the target machine, $\boldsymbol{H}_2 \in \mathbb{R}_{\geq 0}^{\Omega\times L}$ and $\boldsymbol{U}_2 \in \mathbb{R}_{\geq 0}^{L\times T_2}$ are respectively the basis and coefficient matrices that correspond to the operation sound of the target machine as Eq. (31), and $\boldsymbol{G}_2 \in \mathbb{R}_{\geq 0}^{K\times T_2}$ is the coefficient matrix of $\boldsymbol{F}_0$, which corresponds to the background noise.

The anomaly score α is calculated as follows:

$$\alpha = \frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{S}_2'^{(m)} - f_{\mathrm{AE}}\left\{\boldsymbol{S}_2'^{(m)}\middle|\Theta\right\}\right\|_2^2, \tag{32}$$

$$\boldsymbol{S}_2'^{(m)} = \left|\tilde{\boldsymbol{S}}_2^{(m)}\right|^{\cdot\frac{p_{\mathrm{AE}}}{p}}, \tag{33}$$

where $\tilde{\boldsymbol{S}}_2^{(m)} \in \mathbb{R}_{\geq 0}^{\Omega\times T_{\mathrm{AE}}}$ is the spectrogram piece cut out from $\tilde{\boldsymbol{S}}_2$ every time frame number T_{AE}. The threshold value ϕ is determined in advance, and if $\alpha > \phi$, it is judged that an anomaly has occurred on the target machine.

4. Evaluation

In this section, the anomaly detection performance of the proposed system is evaluated. To confirm the validity of the target source extraction process, a system without GGD-NMF (omitting basis training step and GGD-based SS-NMF in the other steps from the proposed system) and the proposed system with various parameter settings were compared.

Table 1 Dataset

	Items	SNR	Time
Training basis matrix $\boldsymbol{F}$	Background noise	-	0.4 h
Training detector	Normal bearing sound + background noise	−24 dB	2.0 h
Testing detection	Normal bearing sound + background noise	−24 dB	5.2 h
	Abnormal bearing sound + background noise	−12 dB	5.2 h
		−18 dB	5.2 h
		−24 dB	5.2 h

4.1 Experimental Conditions

Two types of data, normal and abnormal, were used as samples of the sound to be detected from a dataset published by Case Western Reserve University Bearing Data Center [19]. This dataset consists of vibration signal data from motor bearings acquired using an acceleration sensor and also includes vibration data of normal bearings and abnormal bearings with scratches. The data used as normal were "Normal_2" of "Normal Baseline Data," and the data used as abnormal were "OR007@6_2" of "48k Drive End Bearing Fault Data."

As the background noise, the sound recorded in the equipment room at the Ebina Office of Ricoh Co., Ltd., was used, where several machines such as pumps were operating in the same room. The total recording time was about 5.2 h (18770 s).

These data were mixed, and the experimental data were prepared as shown in Table 1. To train the basis matrix $\boldsymbol{F}_0$, we used 1440-s-long background noise, which was a different time period from the background noise sound for training the detector (AE). The background noise for the detection test includes the time periods used for training the basis matrix and the detector. The volume of background noise is constant in each dataset, namely, when the SNR is changed, only the volume of the bearing is changed. Since the bearing vibration data have a length of only 10 s for both normal and anomaly cases, the background noise data was divided every 10 s, and each divided time period was mixed with the bearing vibration data.

AE for anomaly detection consists of four fully connected layers of the encoder and decoder. The input and output spectrograms are treated as one-dimensional vectors. The number of nodes in each layer of the encoder is halved in order from the input layer, and the decoder has the opposite configuration. Batch normalization [20] is performed on the output of each layer. The ReLU function [21] was used as the activation function. Other experimental conditions are shown in Table 2.

4.2 Experimental Results and Considerations

Figure 2 shows the performance evaluation results of each system. Here, the threshold ϕ for anomaly detection is experimentally determined to be the highest F-measure. In the comparison of the cases without and with NMF preprocessing, the case with preprocessing shows a higher F-measure (with suitable parameter settings). Therefore, it is considered that noise reduction by NMF is an effective preprocessing approach for anomaly detection.

Table 2 Experimental conditions

	Sampling rate	32000 Hz
Window length / shift length		10 ms / 5 ms
	FFT length	512 points (16 ms)
NMF	Number of bases in $\boldsymbol{F}_0$ (K)	256
	Number of bases in $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ (L)	1
AE	Feature	Power spectrogram 256 dimensions × 32 frames ($p_{\mathrm{AE}} = 2, \Omega = 256, T_{\mathrm{AE}} = 32$)
	Structure	Autoencoder
	Nodes	8192-4096-2048-1024 -1024-2048-4096-8192
	Epoch	200

For each case of without NMF, with ISNMF ($p = 0.5$), and with GGD-NMF ($\rho = 4.0$ and $p = 0.5$), the precision–recall curves under each SNR condition are shown in Fig. 3. GGD-NMF has better characteristics than ISNMF, and it is considered that a superior anomaly detection model is generated. Since high performance is obtained in the case of $\rho > p$, it is considered that the mitigation of the low-rank property in NMF modeling is effective for anomaly detection. This is because the mitigation of the low-rank property in the NMF decomposition does not tolerate spectral outliers in the spectrogram, namely, GGD-NMF captures the detailed spectral differences between the normal and abnormal bearing sounds. The mitigation of the low-rank property is derived from the generative model of sub-Gaussian distribution, which corresponds to $\rho > 2$. To the best of our knowledge, no sub-Gaussian-model-based NMF exists other than GGD-NMF. Thus, the proposed system can be considered as the first practical application of sub-Gaussian

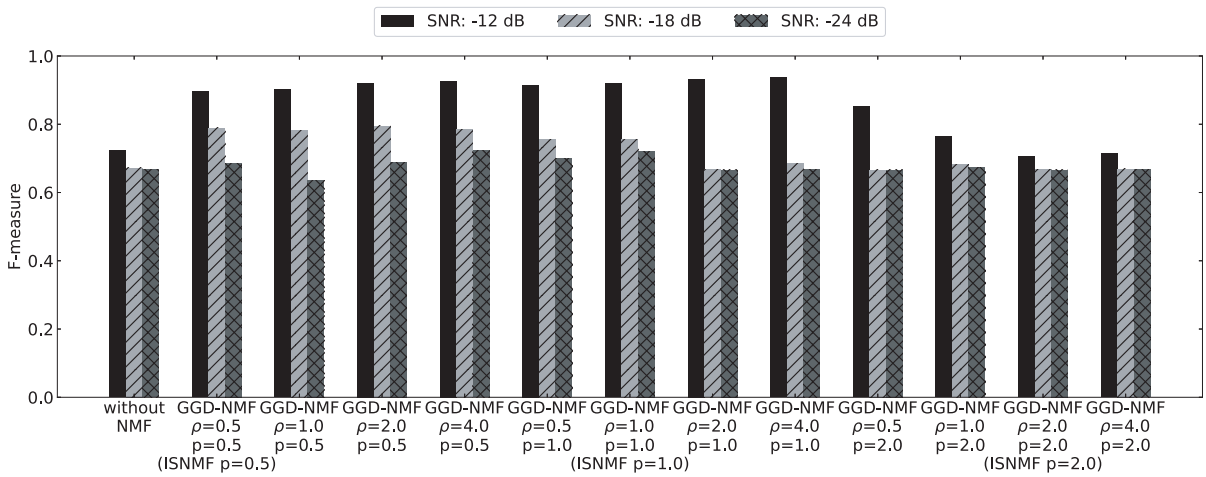
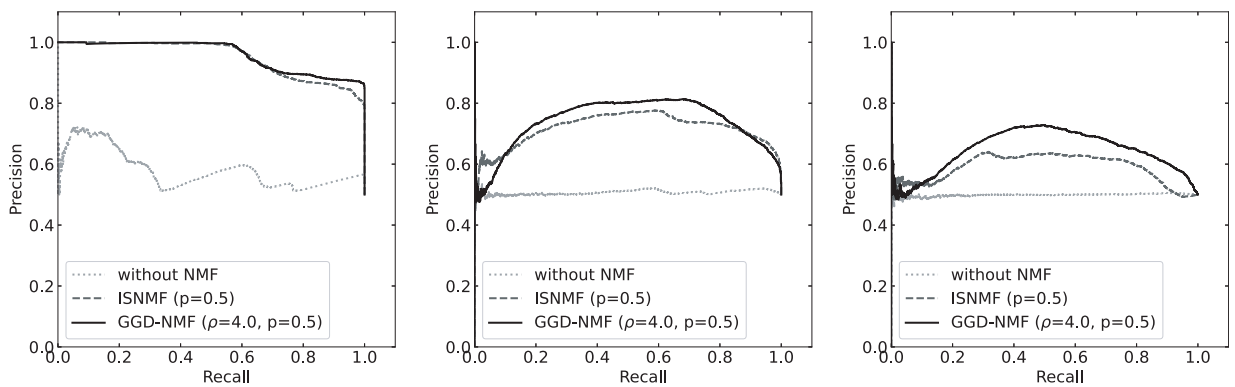


Fig. 2 F-measure values with various GGD-NMF parameters.



(a) SNR: −12 dB (b) SNR: −18 dB (c) SNR: −24 dB

Fig. 3 Precision–Recall curves under each SNR condition.

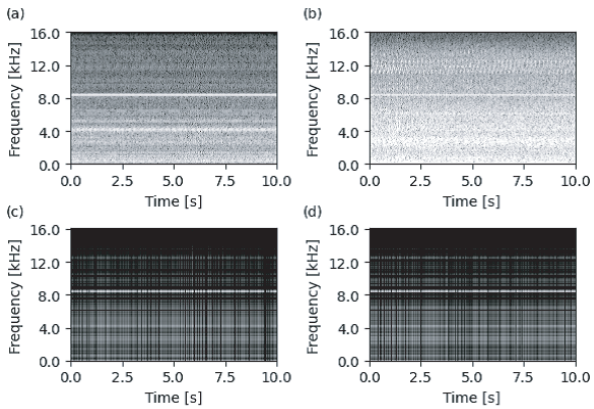


Fig. 4 Examples of power spectrograms of (a) normal bearing vibration, (b) mixed sound of (a) and background noise (SNR: −24 dB), (c) sound extracted by ISNMF ($p = 0.5$) from (b), and (d) sound extracted by GGD-NMF ($\rho = 4.0, p = 0.5$) from (b).

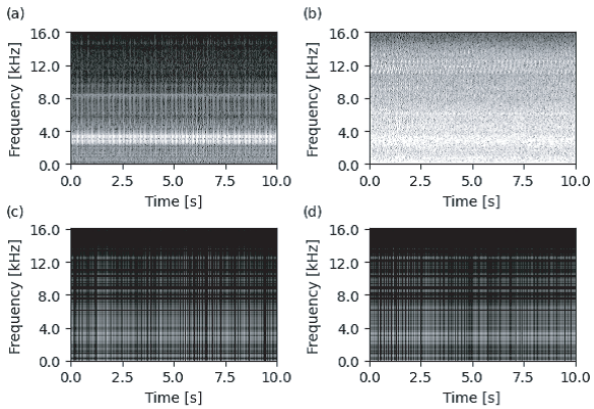


Fig. 5 Examples of power spectrograms of (a) abnormal bearing vibration, (b) mixed sound of (a) and background noise (SNR: −24 dB), (c) sound extracted by ISNMF ($p = 0.5$) from (b), and (d) sound extracted by GGD-NMF ($\rho = 4.0, p = 0.5$) from (b).

NMF, which gives superior performance in an actual task. In addition, the macro model of Eq. (8) with $\rho > 2$ has a versatility as described in Sect. 3.1. This property also contributed to the improvement of the proposed anomaly detection performance.

The effect of GGD-NMF is analyzed from the actual processed signals. Figure 4 shows the examples of normal spectrogram and Fig. 5 shows the examples of abnormal spectrogram. It can be seen that GGD-NMF extracts features of the bearing vibration from the background noise as well as ISNMF does; however, there exists slight difference around 3 kHz between GGD-NMF and ISNMF in Fig. 5(d). To analyze the difference in detail, we next show Figs. 6 and 7 that depict the mean power spectra of the signals extracted by ISNMF ($p = 0.5$) and GGD-NMF ($\rho = 4.0, p = 0.5$), respectively. The larger the difference between the normal and anomaly spectra becomes, the more effective the source extraction process is for anomaly detection. The spectra extracted by GGD-NMF show a greater difference between the normal and abnormal cases in the band of ap-

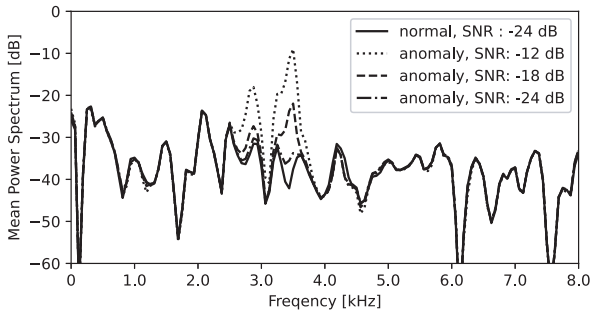


Fig. 6 Mean power spectrum of sounds extracted by ISNMF ($p = 0.5$).

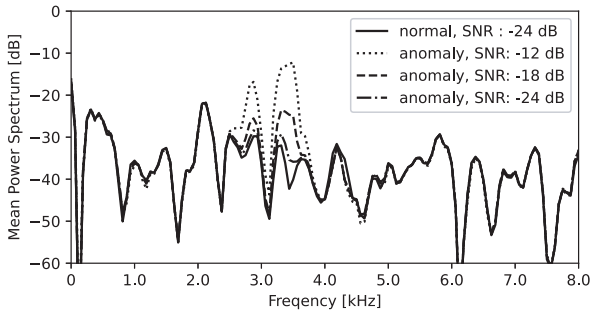


Fig. 7 Mean power spectrum of sounds extracted by GGD-NMF ($\rho = 4.0, p = 0.5$).

proximately 3.0–3.3 kHz than those of ISNMF. This difference is likely to affect the F-measure values. At the 3187.5 Hz frequency bin, the differences between the normal and abnormal mean power spectra extracted by GGD-NMF are greater than those of ISNMF by 5.5 dB (under SNR = −12 dB), 2.4 dB (under SNR = −18 dB), and 0.6 dB (under SNR = −24 dB).

5. Conclusion

In this study, we aimed to improve the performance of the anomaly detection system in a low-SNR environment. For this purpose, we proposed an anomaly detection system based on the outlier detection method with target sound extraction based on GGD-NMF.

From the results of evaluation experiments, it was confirmed that the proposed system can improve the detection performance in low-SNR environments compared with a system without preprocessing. In addition, the mitigation of the low-rank property in NMF resulted in a decomposition more suitable for anomaly detection.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number JP19H01116.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput. Surv., vol.41, no.3, July 2009.

[2] D. Ramotsoela, A. Abu-Mahfouz, and G. Hancke, "A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study," Sensors, vol.18, no.8, p.2491, 2018.
[3] K. Singh and S. Upadhyaya, "Outlier detection: applications and techniques," Int. J. Computer Science Issues, vol.9, no.1, pp.307–323, 2012.
[4] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," SNU Data Mining Center, 2015-2 Special Lecture on IE, 2015.
[5] H. Qiu, J. Lee, J. Lin, and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," J. Sound and Vibration, vol.289, no.4, pp.1066–1090, Feb. 2006.
[6] N. Kitaoka, I. Akahori, and S. Nakagawa, "Speech recognition under noisy environments using spectral subtraction with smoothing of time direction and real-time cepstral mean normalization," Proc. International Workshop on Hands-Free Speech Communication, pp.159–162, 2001.
[7] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401, no.6755, p.788, Oct. 1999.
[8] X. lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising auto-encoder," Proc. Interspeech, pp.436–440, 2013.
[9] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio, Speech, Language Process., vol.26, no.10, pp.1702–1726, Oct. 2018.
[10] H. Shao, H. Jiang, F. Wang, and H. Zhao, "An enhancement deep feature fusion method for rotating machinery fault diagnosis," Knowledge-Based Systems, vol.119, pp.200–220, March 2017.
[11] D. Kitamura, "Nonnegative matrix factorization based on complex generative model," Acoustical Science and Technology, vol.40, no.3, pp.155–161, 2019.
[12] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," Proc. the 7th Int. Conf. Independent Component Analysis and Signal Separation, pp.414–421, 2007.
[13] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised non-negative matrix factorization with orthogonality and maximum-divergence penalties," IEICE Trans. Fundamentals, vol.97-A, no.5, pp.1113–1118, May 2014.
[14] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural Computation, vol.21, no.3, pp.793–830, March 2009.
[15] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems, vol.13, pp.556–562, 2000.
[16] S. Mogami, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, and N. Ono, "Independent low-rank matrix analysis based on time-variant sub-gaussian source model for determined blind source separation," IEEE/ACM Trans. Audio, Speech, Language Process., vol.28, pp.503–518, Jan. 2019.
[17] K. Kamo, Y. Kubo, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Joint-diagonalizability-constrained multichannel nonnegative matrix factorization based on multivariate complex sub-gaussian distribution," Proc. EUSIPCO2020, 2020. (arXiv preprint arXiv:2007.00416).
[18] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," EURASIP J. Adv. Signal Process., vol.2018, no.1, Article No. 28, May 2018.
[19] Case Western Reserve University, "Bearing Data Center," http://csegroups.case.edu/bearingdatacenter/home, accessed May. 12. 2017.
[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.
[21] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," Proc. the Fourteenth Int. Conf. Artificial Intelligence and Statistics, pp.315–323, 2011.

Akihito Aiba received the B.E. and M.E. degrees from Tohoku University in 2008 and 2010, respectively. He joined Ricoh Company, Ltd. in 2010. His research interests include acoustic signal processing and machine learning. He is a member of the Acoustical Society of Japan.

Minoru Yoshida received the M.E. degree from Tokyo Denki University in acoustical digital signal processing in 1993. He joined Ricoh Company, Ltd. in 2015. His research interests are digital signal processing for outlier detection.

Daichi Kitamura received the Ph.D. degree from SOKENDAI, Hayama, Japan. He joined the University of Tokyo, Tokyo, Japan, in 2017 as a Research Associate, and he moved to the National Institute of Technology, Kagawa College, Takamatsu, Japan, as an Assistant Professor, in 2018. His research interests include audio source separation, statistical signal processing, and machine learning. He was the recipient of the Awaya Prize Young Researcher Award from The Acoustical Society of Japan (ASJ) in 2015, Ikushi Prize from Japan Society for the Promotion of Science in 2017, Best Paper Award from IEEE Signal Processing Society Japan in 2017, and Itakura Prize Innovative Young Researcher Award from ASJ in 2018.

Shinnosuke Takamichi received the B.E. degree from Nagaoka University of Technology, Nagaoka, Japan, in 2011, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, in 2013 and 2016, respectively. He is currently an Assistant Professor at The University of Tokyo. He has received more than ten paper/achievement awards including the 3rd IEEE Signal Processing Society Japan Young Author Best Paper Award. He is a member of ASJ, IPSJ, ISCA, and IEEE SPS.

Hiroshi Saruwatari Hiroshi Saruwatari received the B.E., M.E., and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM Intelligent Systems Laboratory, Tokyo, Japan, in 1993, and the Nara Institute of Science andTechnology, Ikoma, Japan, in 2000. Since 2014, he has been a Professor with the University of Tokyo, Tokyo. His research interests include statistical audio signal processing, blind source separation, and speech enhancement. He was the recipient of the paper awards from IEICE in 2001 and 2006, fromTAF in 2004, 2009, 2012, and 2018, from IEEE-IROS2005 in 2006, and from APSIPA in 2013 and 2018. He was also the recipient of the DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, The Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hoko-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ.