Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma

Yuma Koizumi, *Member, IEEE*, Shoichiro Saito, *Member, IEEE*, Hisashi Uematsu, Yuta Kawachi, and Noboru Harada, *Senior Member, IEEE*

***Abstract*—This paper proposes a novel optimization principle and its implementation for unsupervised anomaly detection in sound (ADS) using an autoencoder (AE). The goal of the unsupervised-ADS is to detect unknown anomalous sounds without training data of anomalous sounds. The use of an AE as a normal model is a state-of-the-art technique for the unsupervised-ADS. To decrease the false positive rate (FPR), the AE is trained to minimize the reconstruction error of normal sounds, and the anomaly score is calculated as the reconstruction error of the observed sound. Unfortunately, since this training procedure does not take into account the anomaly score for anomalous sounds, the true positive rate (TPR) does not necessarily increase. In this study, we define an objective function based on the Neyman–Pearson lemma by considering the ADS as a statistical hypothesis test. The proposed objective function trains the AE to maximize the TPR under an arbitrary low FPR condition. To calculate the TPR in the objective function, we consider that the set of anomalous sounds is the complementary set of normal sounds and simulate anomalous sounds by using a rejection sampling algorithm. Through experiments using synthetic data, we found that the proposed method improved the performance measures of the ADS under low FPR conditions. In addition, we confirmed that the proposed method could detect anomalous sounds in real environments.**

***Index Terms*—Anomaly detection in sound, Neyman-Pearson lemma, deep learning, and autoencoder.**

I. Introduction

ANOMALY detection in sound (ADS) has received much attention. Since anomalous sounds might indicate symptoms of mistakes or malicious activities, their prompt detection can possibly prevent such problems. In particular, ADS has been used for various purposes including audio surveillance [1]–[4], animal husbandry [5], [6], product inspection, and predictive maintenance [7], [8]. For the last application, since anomalous sounds might indicate a fault in a piece of machinery, prompt detection of anomalies would decrease the number of defective product and/or prevent propagation of damage. In this study, we investigated ADS for industrial equipment by focusing on machine-operating sounds.

ADS tasks can be broadly divided into supervised-ADS and unsupervised-ADS. The difference between the two categories is in the definition of anomalies. Supervised-ADS is the task of detecting "*defined*" anomalous sounds such as gunshots or screams [2], and it is a kind of rare sound event detection (SED) [9]–[11]. Since the anomalies are defined, we can collect a dataset of the target anomalous sounds even though the anomalies are rarer than normal sounds. Thus, the ADS system can be trained using a supervised method that is used in various SED tasks of the "Detection and Classification of Acoustic Scenes and Events challenge" (DCASE) such as audio scene classification [12], [13], sound event detection [14], [15], and audio tagging [16]. On the other hand, unsupervised-ADS [17]–[19] is the task of detecting "*unknown*" anomalous sounds that have not been observed. In the case of real-world factories, from the view of the development cost, it is impracticable to deliberately be damaged the expensive target machine. In addition, actual anomalous sounds occur rarely and have high variability. Therefore, it is impossible to collect an exhaustive set of anomalous sounds and need to detect anomalous sounds for which training data does not exist. From this reason, the task is often tackled as the one-class unsupervised classification problem [17]–[19]. This point is one of the major differences in premise between the DCASE tasks and ADS for industrial equipment. Thus, in this study, we aim to detect unknown anomalous sounds based on an unsupervised approach.

In unsupervised anomaly detection, "*anomaly*" is defined as the patterns in data that do not conform to expected "*normal*" behavior [19]. Namely, the universal set consists of only the normal and the anomaly, and the anomaly is the complement to the normal set. More intuitively, the universal set is various machine sounds including many types of machines, the normal set is one specific type of various machine sound, and the anomaly set is all other types of machine sounds. Therefore, a typical way of unsupervised-ADS is the use of the outlier-detection technique. Here, the deviation between a normal model and an observed sound is calculated; the deviation is often called the "*anomaly score*". The normal model indicates the notion of normal behavior which is trained from training data of normal sounds. The observed sound is identified as an anomalous one when the anomaly score is higher than a pre-defined threshold value. Namely, the anomalous sounds are defined as the sounds that do not exist in training data of normal sounds.

Manuscript received April 24, 2018; revised August 21, 2018 and October 4, 2018; accepted October 16, 2018. Date of publication October 22, 2018; date of current version November 28, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Augusto Sarti. *(Corresponding author: Yuma Koizumi.)*

The authors are with the NTT Media Intelligence Laboratories, NTT Corporation, Tokyo 180-8585, Japan (e-mail: koizumi.yuma@ieee.org; saito.shoichiro@lab.ntt.co.jp; uematsu.hisashi@lab.ntt.co.jp; kawachi.yuta@lab.ntt.co.jp; noboru.harada@lab.ntt.co.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2877258

2329-9290 © 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

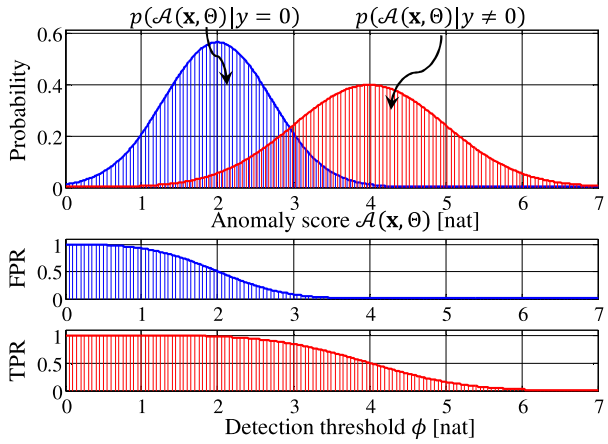


Fig. 1. Trade-off relationship between anomaly score, true positive rate (TPR) and false positive rate (FPR). The top figure shows PDFs of anomaly scores for normal sounds (blue line) and anomalous sounds (red line). The bottom figures show the FPR and TPR with respect to the threshold. When these PDFs overlap, a small threshold leads to a large TPR and FPR, and a large threshold leads to a small TPR and FPR.

To train the normal model, it is necessary to define the optimality of the anomaly score. One of the popular performance measurements of ADS is to measure both the true positive rate (TPR) and false positive rate (FPR). The TPR is the proportion of anomalies that are correctly identified, and the FPR is the proportion of normal sounds that are incorrectly identified as anomalies. To improve the performance of ADS, we need to increase TPR and decrease FPR simultaneously. However, these metrics are related to the threshold value and have a trade-off relationship, as shown in Fig. 1. When the PDFs of the anomaly scores of normal and anomalous sounds overlap, false detections cannot be avoided regardless of any threshold. Thus, to increase TPR and decreases FPR simultaneously, we need to train the normal model to reduce the overlap area. More intuitively, it is essential to provide small anomaly scores for normal sounds and large anomaly scores for anomalous sounds. In addition, if an ADS system gives a false alert frequently, we cannot trust it, just as "*the boy who cried wolf*" cannot be trusted. Therefore, it is especially important to increase TPR under a low FPR condition in a practical situation.

The early studies used various statistical models to calculate the anomaly score, such as the Gaussian mixture model (GMM) [3], [8] and support vector machine (SVM) [4]. The recent literature calculates the anomaly score through the use of deep neural networks (DNN) such as the autoencoder (AE) [20]–[23] and variational AE (VAE) [24], [25]. In the case of the AE, one is trained to minimize the reconstruction error of the normal training data, and the anomaly score is calculated as the reconstruction error of the observed sound. Thus, the AE provides small anomaly scores for normal sounds. However, it gives no guarantee to increase anomaly scores for anomalous sounds. Indeed, if the AE is generalized, the anomalous sounds will also be reconstructed and the anomaly score of anomalous sound will be small. Therefore, to increase TPR and decrease FPR simultaneously, the objective function should be modified.

Another strategy for unsupervised-ADS is the use of a generative adversarial network (GAN) [26], [27]. GANs have been used to detect anomalies in medical images [28]. In this strategy, a generator simulates "fake" normal data, and a discriminator identifies whether the input data is a real normal data or not. Therefore, the discriminator can be trained to increase TPR for fake normal data and decrease FPR for true normal data simultaneously. However, since the generator is trained to make normal data, if it perfectly generates normal sounds, the anomaly score of normal sounds and FPR will increase. Therefore, it is necessary to build an algorithm to simulate "non-normal" sounds.

In this study, we propose a novel optimization principle and its implementation for ADS using AE. By considering an outlier-detection-based ADS as a statistical hypothesis test, we define optimality as an objective function based on the Neyman-Pearson lemma [29]. The objective function works to increase TPR under an arbitrary low FPR condition. A problem in calculating TPR is the simulation of anomalous sound data. Here, we explicitly define the set of anomalous sounds to be the complement to the set of normal sounds and simulate anomalous sounds by using a rejection sampling algorithm.

A preliminary version of this work is presented in [8]. The previous study utilized a DNN as a feature extractor, and the anomaly score was calculated using the negative-log-likelihood of a GMM trained from normal data. Thus, although the DNN was trained to maximize the objective function based on the Neyman-Pearson lemma, the normal model did not guarantee to increase TPR and decrease FPR. In this study, end-to-end training is achieved by using an AE as the normal model and both the feature extractor and the normal model are trained to increase TPR and decrease FPR.

The rest of this paper is organized as follows. Section II briefly introduces outlier-detection-based ADS and its implementation using an AE. Section III describes the proposed training method and the details of the implementation. After reporting the results of objective experiments using synthetic data and verification experiments in real environments in Section IV, we conclude this paper in Section V. The mathematical symbols are listed in Appendix A.

II. Conventional Method

A. Identification of Anomalous Sound Based on Outlier Detection

ADS is an identification problem of determining whether the sound emitted from a target is a normal sound or an anomalous one. In this section, we briefly introduce the procedure of unsupervised-ADS.

First, an anomaly score $\mathcal{A}(\mathbf{x}_\tau, \Theta)$ is calculated using a normal model. Here, $\mathbf{x}_\tau \in \mathbb{R}^Q$ is an input vector calculated from the observed sound indexed on $\tau \in \{1, 2, ..., T\}$ for time, and Θ is the set of parameters of the normal model. In many of the previous studies, $\mathbf{x}_\tau$ was composed of hand-crafted acoustic features such as mel-frequency cepstrum coefficients (MFCCs) [1]–[3], and the normal model was often constructed with a PDF of normal sounds. Accordingly, the anomaly score can be

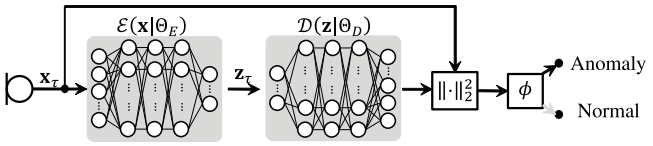


Fig. 2. Anomaly detection procedure using autoencoder. The input vector is compressed and reconstructed by two networks $\mathcal{E}$ and $\mathcal{D}$, respectively. Since $\mathcal{E}$ and $\mathcal{D}$ are trained to minimize reconstruction error of normal sounds, the reconstruction error would be small if $\mathbf{x}_\tau$ is normal. Thus, the anomaly score is calculated as a reconstruction error, and when the error exceeds a pre-defined threshold ϕ, the observation is identified as anomalous.

calculated as

$$\mathcal{A}(\mathbf{x}_\tau, \Theta) = -\ln p(\mathbf{x}_\tau \mid \Theta, y = 0), \tag{1}$$

where y denotes the state, $y = 0$ is normal, and $y \neq 0$ is not normal, i.e. anomalous. $p(\mathbf{x}|\Theta, y = 0)$ is a normal model such as a GMM [8]. $\mathbf{x}_\tau$ is determined to be anomalous when the anomaly score exceeds a pre-defined threshold value ϕ:

$$\mathcal{H}(\mathbf{x}_\tau, \Theta, \phi) = \begin{cases} 0 \text{ (Normal)} & \mathcal{A}(\mathbf{x}_\tau, \Theta) \leq \phi \\ 1 \text{ (Anomaly)} & \mathcal{A}(\mathbf{x}_\tau, \Theta) > \phi \end{cases}. \tag{2}$$

One of the performance measures of ADS consists of the pair of TPR and FPR. The TPR and FPR can be calculated as expectations of $\mathcal{H}(\mathbf{x}, \Theta, \phi)$ with respect to anomalous and normal sounds, respectively:

$$\text{TPR}(\Theta, \phi) = \mathbb{E}\left[\mathcal{H}(\mathbf{x}, \Theta, \phi)\right]_{\mathbf{x}|y\neq 0}, \tag{3}$$

$$\text{FPR}(\Theta, \phi) = \mathbb{E}\left[\mathcal{H}(\mathbf{x}, \Theta, \phi)\right]_{\mathbf{x}|y= 0}, \tag{4}$$

where $\mathbb{E}[\cdot]_x$ denotes the expectation with respect to x. These metrics are related to ϕ and have a trade-off relationship as shown in Fig. 1. The top figure shows the PDFs of anomaly scores for normal sounds $p(\mathcal{A}(\mathbf{x}_\tau, \Theta)|y = 0)$ and anomalous sounds $p(\mathcal{A}(\mathbf{x}_\tau, \Theta)|y \neq 0)$. The bottom figures show the FPR and TPR with respect to ϕ. When these PDFs overlap, false detections, i.e. false-positive and/or false-negative, cannot be avoided regardless of any ϕ. In addition, the false detections increase as the overlap area gets wider. Therefore, to increase TPR and decrease FPR simultaneously, it is necessary to train Θ so that the anomaly score is small for normal sounds and large for anomalous sounds. More precisely, we need to train Θ to reduce the overlap area.

B. Unsupervised-ADS Using an Autoencoder

Recently, deep learning has been used to construct a normal model. Several studies on deep-learning-based unsupervised-ADS have used an autoencoder (AE) [20]–[23]. This section briefly describes unsupervised-ADS using an AE (see Fig. 2).

The goal of using an AE is to learn an efficient representation of the input vector by using two neural networks $\mathcal{E}$ and $\mathcal{D}$, which are called the encoder and decoder, respectively. First, the input vector $\mathbf{x}$ is converted into a latent vector $\mathbf{z} \in \mathbb{R}^{\mathsf{R}}$ by $\mathcal{E}$. Then, an input vector is reconstructed from $\mathbf{z}$ by $\mathcal{D}$. These processes are expressed as

$$\mathbf{z} = \mathcal{E}(\mathbf{x} \mid \Theta_E), \tag{5}$$

$$\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z} \mid \Theta_D). \tag{6}$$

The parameters of both neural networks $\Theta = \{\Theta_E, \Theta_D\}$ are trained to minimize the reconstruction error:

$$\mathcal{J}^{\text{AE}}(\Theta_E, \Theta_D) = \mathbb{E}\left[\|\mathbf{x} - \mathcal{D}(\mathcal{E}(\mathbf{x} \mid \Theta_E) \mid \Theta_D)\|_2^2\right]_{\mathbf{x}}. \tag{7}$$

In ADS using an AE, the anomaly score is the reconstruction error of the observed sound, which is calculated as

$$\mathcal{A}(\mathbf{x}_\tau, \Theta) := \|\mathbf{x}_\tau - \mathcal{D}(\mathcal{E}(\mathbf{x}_\tau \mid \Theta_E) \mid \Theta_D)\|_2^2. \tag{8}$$

To train the normal model to provide small anomaly scores for normal sounds, the AE is trained to minimize the average reconstruction error of normal sound,

$$\mathcal{J}^{\text{AE}}(\Theta_E, \Theta_D) = \frac{1}{N^{(u)}} \sum_{n=1}^{N^{(u)}} \mathcal{A}(\mathbf{x}_n^{(u)}, \Theta), \tag{9}$$

where $\mathbf{x}_n^{(u)}$ is the n-th training data of normal sound and $N^{(u)}$ is the number of training samples of normal sound. This objective function works to decrease the anomaly score of normal sounds. However, there is no guarantee of increasing anomaly scores for anomalous sounds. Indeed, if the AE is generalized, the anomalous sounds will also be reconstructed and the anomaly score of anomalous sounds will be also small. Therefore, (9) does not ensure that false detections are reduced and the accuracy of ADS is improved; thus, it would be better to modify the objective function.

III. Proposed Method

We will begin by defining an objective function that builds upon the Neyman-Pearson lemma in Section III-A. Then, we will describe the rejection sampling algorithm for simulating anomalous sound used for calculating TPR in Section III-B. After that, the overall training and detection procedure of the proposed method will be summarized in Section III-C and Section III-D. As a modified implementation of proposed method, we extend the proposed method to an area under the receiver operating characteristic curve (AUC) maximization in Section III-E.

A. Objective Function for Anomaly Detection Based on the Neyman-Pearson Lemma

From (1) and (2), an anomalous sound satisfies the following inequality:

$$p(\mathbf{x} \mid \Theta, y = 0) < \exp(-\phi). \tag{10}$$

Since ϕ is assumed to be sufficiently large to avoid false positives, an anomalous sound can be defined as "a sound which cannot be regarded as a sample of the normal model." Thus, we can regard outlier-detection-based ADS as a statistical hypothesis test. In other words, the observed sound is identified as anomalous when the following null hypothesis is rejected.

Null hypotheses: $\mathbf{x}$ is a sample of the normal model $p(\mathbf{x} \mid \Theta, y = 0)$.

The Neyman-Pearson lemma [29] states the condition for $\mathcal{A}(\mathbf{x}, \Theta)$ that achieves the *most powerful test* between two simple hypotheses. According to it, the most powerful test has the greatest detection power among all possible tests of a given FPR [30]. More simply, the most powerful test maximizes the TPR under the constraint that the FPR equals ρ, i.e.,

$$\text{maximize TPR}(\Theta, \phi), \text{ subject to FPR}(\Theta, \phi) = \rho.$$

Since the FPR can be controlled by manipulating ϕ, we define ϕ_ρ as satisfying $\text{FPR}(\Theta, \phi_\rho) = \rho$. Accordingly, the objective function to obtain the most powerful test function can be defined as the one that maximizes $\text{TPR}(\Theta, \phi_\rho)$ with respect to Θ. However, since the FPR is also a function of Θ, it may become large when focusing only on TPR. To maximize the TPR and minimize the FPR simultaneously, we train Θ to maximize the following objective function,

$$\mathcal{J}^{\text{NP}}(\Theta) = \text{TPR}(\Theta, \phi_\rho) - \text{FPR}(\Theta, \phi_\rho), \tag{11}$$

where the superscript "NP" is an abbreviation of "Neyman-Pearson". Since the proposed objective function directly increases TPR and decreases FPR, Θ can be trained to provide a small anomaly score for normal sounds and a large anomaly score for anomalous sounds.

There are two problems when it comes to training Θ_E and Θ_D to maximize (11). The first problem is the calculation of TPR. The TPR and FPR are the expectations of $\mathcal{H}(\mathbf{x}, \Theta, \phi)$, and in most practical cases, the expectation is approximated as an average over the training data. Thus, to calculate TPR and FPR, we need to collect enough normal and anomalous sound data for the average to be an accurate approximation of the expectation. However, since anomalous sounds occur rarely and have high variability, this condition is difficult to satisfy. In Section III-B, to calculate TPR, we consider "anomaly" to mean "not normal" and simulate anomalous sounds by using a sampling algorithm. The second problem is the determination of the threshold ϕ_ρ. In a parametric hypothesis test such as a t-test, the threshold at which FPR equals ρ can be analytically calculated. However, DNN is a non-parametric statistical model; thus, the threshold ϕ_ρ can not be analytically calculated. In Section III-C, we numerically calculate ϕ_ρ as the $\lfloor \rho M \rfloor$-th value of the sorted anomaly scores of M normal sounds, where $\lfloor \cdot \rfloor$ is the flooring function.

B. Anomalous Sound Simulation Using an Autoencoder

In accordance with (10), anomalous sounds emitted from the target machine are different from normal ones. Thus, we consider the set of normal sounds to be a subset of various machine sounds, and the set of anomalous sounds to be its complement. Then, we use rejection sampling to simulate anomalous sounds; namely, a sound is sampled from various machine-sound PDFs, and it is accepted as an anomalous sound when its anomaly score is high. However, since the PDF of various machine sounds in the input vector domain $p(\mathbf{x})$ may have a complex form, the PDF cannot be written in an analytical form and the sampling algorithm would become complex. Inspired by the strategy of VAE, we can avoid this problem by training $\mathcal{E}$ so that the PDF of various latent vectors $p(\mathbf{z})$ is mapped to a PDF whose samples

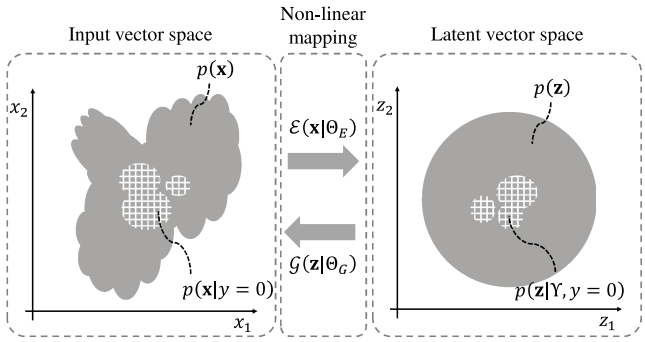


Fig. 3. Concept of PDFs of normal, various, and anomalous sounds using two neural networks. The PDF of normal sounds (i.e. meshed area) is a subset of the PDF of various sounds (i.e. gray area), and the PDF of anomalous sounds is expressed as complement of the PDF of normal sounds (i.e. inside the gray area and outside the meshed area). $\mathbf{x}$ is mapped to $\mathbf{z}$ by $\mathcal{E}$, and $\mathbf{z}$ is reconstructed to $\tilde{\mathbf{x}}$ by $\mathcal{G}$. Here, $\mathcal{E}$ and $\mathcal{G}$ are trained to satisfy $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}_{\mathsf{R}}, \boldsymbol{I}_{\mathsf{R}})$ and $\mathbf{x} = \tilde{\mathbf{x}}$, respectively. The PDF of the latent vector of normal sounds is modeled using a GMM $p(\mathbf{z} \mid \Upsilon, y = 0)$ given by (13).

can be generated by a pseudorandom number generator from a uniform distribution and its variable conversion. Then, the latent vectors of anomalous sounds $\mathbf{z}^{(a)}$ are sampled using the rejection sampling algorithm, and the input vectors of anomalous sounds $\mathbf{x}^{(a)}$ are reconstructed using a third neural network $\mathcal{G}$,

$$\mathbf{x}^{(a)} = \mathcal{G}(\mathbf{z}^{(a)} \mid \Theta_G), \tag{12}$$

where Θ_G is the parameter of $\mathcal{G}$. Hereafter, we call $\mathcal{G}$ the generator. Although there is no constraint on the architecture of $\mathcal{G}$, we will use the same architecture for $\mathcal{D}$ and $\mathcal{G}$. In addition, to simply generate and reject a candidate latent vector, we use two constraints to train Θ_E and Θ_G, and model the PDF of normal latent vectors using the GMM as

$$p(\mathbf{z} \mid \Upsilon, y = 0) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{13}$$

where $\Upsilon = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid k = 1, ..., K\}$, K is the number of mixtures, and w_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ are respectively the weight, mean vector, and covariance matrix of the k-th Gaussian. The concepts of these PDFs are shown in Fig. 3, and the procedure of anomalous sound simulation is summarized in **Algorithm 1** and Fig. 4.

First, we describe the two constraints for training Θ_E and Θ_G. For algorithmic efficiency, $p(\mathbf{z})$ should be generated with a low computational cost. As an implementation of $p(\mathbf{z})$, we use the normalized Gaussian distribution, because its samples can be generated by a pseudorandom number generator such as the Mersenne-Twister. Thus, for training Θ_E and Θ_G, we use the first constraint so that $\mathbf{z}$ of the various machine sounds follows a normalized Gaussian distribution. To satisfy the first constraint, we train Θ_E to minimize the following Kullback-Leibler divergence (KLD):

$$\begin{aligned}\mathcal{J}^{\text{KL}}(\Theta_E) &= D\left(\mathcal{N}(\mathbf{z} \mid \mathbf{0}_{\mathsf{R}}, \boldsymbol{I}_{\mathsf{R}}) \| \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_{\mathcal{V}}, \boldsymbol{\Sigma}_{\mathcal{V}})\right), \\ &= \frac{1}{2}\left[\ln|\boldsymbol{\Sigma}_{\mathcal{V}}| + \text{tr}\left\{\boldsymbol{\Sigma}_{\mathcal{V}}^{-1}\right\} + \boldsymbol{\mu}_{\mathcal{V}}^{\top}\boldsymbol{\Sigma}_{\mathcal{V}}^{-1}\boldsymbol{\mu}_{\mathcal{V}} - \mathsf{R}\right],\end{aligned} \tag{14}$$

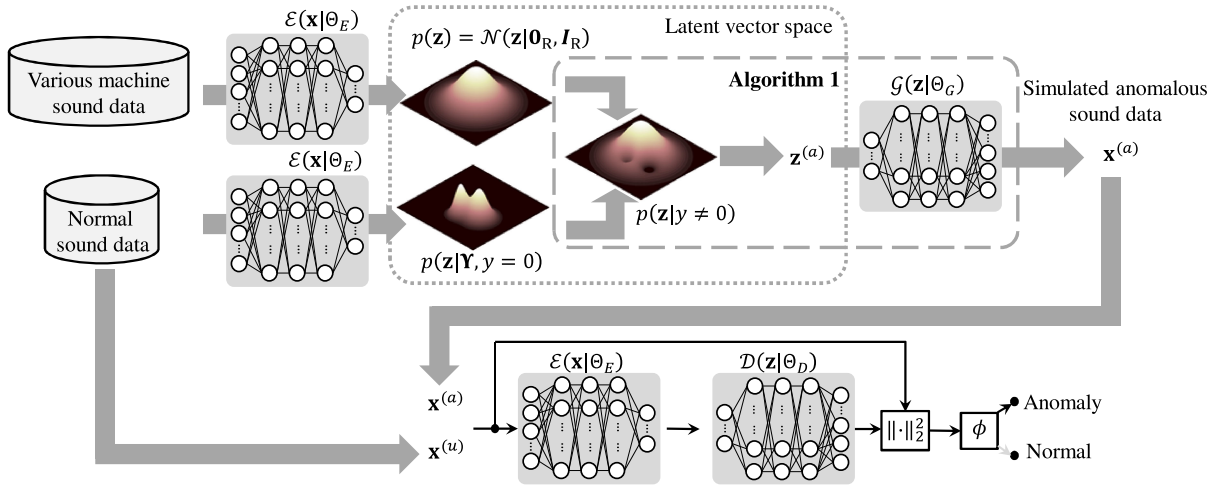


Fig. 4. Procedure of anomalous sound simulation using autoencoder.

Algorithm 1: Simulation Algorithm of Anomalous Sound in Latent Vector Space.

Input: Generator $\mathcal{G}$, GMM $p(\mathbf{z} \mid \Upsilon, y = 0)$ and ϕ_z
$\ell \leftarrow -\infty$
while $\ell \leq \phi_z$ **do**
  Draw $\tilde{\mathbf{z}}$ from $\mathcal{N}(\mathbf{z}|\mathbf{0}_\mathsf{R}, \boldsymbol{I}_\mathsf{R})$
  Evaluate $\ell \leftarrow -\ln p(\tilde{\mathbf{z}} \mid \Upsilon, y = 0)$
end while
$\mathbf{z}^{(a)} \leftarrow \tilde{\mathbf{z}}$
Generate anomalous sound by $\mathbf{x}^{(a)} = \mathcal{G}(\mathbf{z}^{(a)} \mid \Theta_G)$
Output: $\mathbf{x}^{(a)}$

where the superscript "KL" is an abbreviation of "Kullback-Leibler", $\mathrm{tr}\{\cdot\}$ denotes the trace of a matrix, $\top$ denotes transposition, $\mathbf{0}_\mathsf{R}$ and $\boldsymbol{I}_\mathsf{R}$ are respectively the zero vector and unit matrix with dimension R, and $\boldsymbol{\mu}_\mathcal{Y}$ and $\boldsymbol{\Sigma}_\mathcal{Y}$ are respectively the mean vector and covariance matrix calculated from $\mathbf{z}$ of the various machine sounds. To generate anomalous sounds from (12), $\mathcal{G}$ needs to reconstruct various machine sounds, as $\mathbf{x}^{(v)} = \mathcal{G}(\mathcal{E}(\mathbf{x}^{(v)} \mid \Theta_E) \mid \Theta_G)$. Thus, as a second constraint, we train Θ_E and Θ_G to minimize the reconstruction error (7) calculated on the various machine sounds.

Next, we describe the GMM that models the PDF of the normal latent vectors. To reject a candidate $\tilde{\mathbf{z}}$ which seems to be $\mathbf{z}$ of a normal sound, we need to calculate the probability that the candidate is a normal one. To calculate the probability, we need to model $p(\mathbf{z} \mid y = 0)$. Since there is no constraint on the form of $p(\mathbf{z} \mid y = 0)$ in the training procedure of Θ_E, $p(\mathbf{z} \mid y = 0)$ might have a complex form. For simplicity, we use a GMM expressed as (13).

C. Detailed Description of Training Procedure

Here, we describe the details of the training procedure shown in Fig 5. The training procedure consists in three steps. Hereafter, we call the proposed method using this training procedure NP-PROP. The algorithm inputs are training data constructed from normal sounds and various machine sounds, and the outputs are Θ_E and Θ_D. Moreover, $\mathbf{x}_n^{(v)}$ and $\mathbf{x}_n^{(u)}$ respectively denote the n-th training samples of minibatches of various and normal machine sounds, and M is the number of samples included in a minibatch.

First, Θ_E and Θ_G are trained to simulate anomalous sounds. A minibatch of various machine sounds is randomly selected from the training dataset of various machine sounds. Next, its latent vectors are calculated as $\mathbf{z}_n^{(v)} \leftarrow \mathcal{E}(\mathbf{x}_n^{(v)}|\Theta_E)$. Then, the parameters of the Gaussian distribution of the minibatch are calculated as

$$\boldsymbol{\mu}_\mathcal{Y} = \frac{1}{M}\sum_{n=1}^{M} \mathbf{z}_n^{(v)}, \tag{15}$$

$$\boldsymbol{\Sigma}_\mathcal{Y} = \frac{1}{M}\sum_{n=1}^{M} \left(\mathbf{z}_n^{(v)} - \boldsymbol{\mu}_\mathcal{Y}\right)\left(\mathbf{z}_n^{(v)} - \boldsymbol{\mu}_\mathcal{Y}\right)^\top. \tag{16}$$

Finally, to minimize the KLD and the reconstruction error of various sounds, the objective function is calculated as

$$\mathcal{J}^{\mathrm{KR}}(\Theta) = \mathcal{J}^{\mathrm{KL}}(\Theta_E) + \sum_{n=1}^{M} \left\| \mathbf{x}_n^{(v)} - \mathcal{G}\left(\mathcal{E}\left(\mathbf{x}_n^{(v)} \mid \Theta_E\right) \mid \Theta_G\right)\right\|_2^2, \tag{17}$$

where the superscript "KR" is an abbreviation of "KLD and reconstruction", and Θ_E and Θ_G are updated by gradient descent to minimize $\mathcal{J}^{\mathrm{KR}}(\Theta)$:

$$\Theta_E \leftarrow \Theta_E - \lambda \nabla_{\Theta_E} \mathcal{J}^{\mathrm{KR}}(\Theta), \tag{18}$$

$$\Theta_G \leftarrow \Theta_G - \lambda \nabla_{\Theta_G} \mathcal{J}^{\mathrm{KR}}(\Theta), \tag{19}$$

where λ is the step size.

Second, Θ_E and Θ_D are trained to maximize the objective function. A minibatch of normal sounds $\mathbf{x}^{(u)}$ is randomly selected from the training dataset of normal sounds, and a minibatch of anomalous sounds $\mathbf{x}^{(a)}$ is simulated using **Algorithm 1**. Here, since DNN is not a parametric PDF, the threshold ϕ_ρ that satisfies $\mathrm{FPR}(\Theta, \phi_\rho) = \rho$ cannot be analytically calculated.

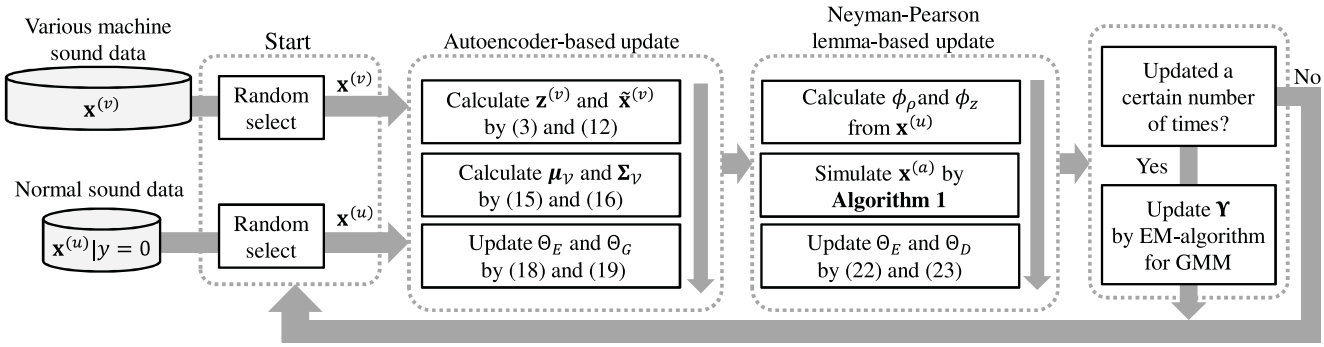


Fig. 5. Training procedure of the proposed method.

Thus, in this study, we approximately calculate ϕ_ρ by sorting the anomaly scores of normal sounds in the minibatch $\mathbf{x}^{(u)}$. First, $\mathcal{A}(\mathbf{x}^{(u)}, \Theta)$ and $-\ln(\mathbf{z}^{(u)} \mid \Upsilon, y=0)$ are calculated, and ϕ_ρ and ϕ_z are set as the $\lfloor \rho M \rfloor$-th value of the sorted $\mathcal{A}(\mathbf{x}^{(u)}, \Theta)$ and $-\ln(\mathbf{z}^{(u)} \mid \Upsilon, y=0)$ in descending order, respectively. Then, the TPR and FPR are approximately evaluated as

$$\mathrm{TPR}(\Theta, \phi_\rho) \approx \frac{1}{M} \sum_{n=1}^{M} \mathrm{sigmoid}\left(\mathcal{A}\left(\mathbf{x}_n^{(a)}, \Theta\right) - \phi_\rho\right), \quad (20)$$

$$\mathrm{FPR}(\Theta, \phi_\rho) \approx \frac{1}{M} \sum_{n=1}^{M} \mathrm{sigmoid}\left(\mathcal{A}\left(\mathbf{x}_n^{(u)}, \Theta\right) - \phi_\rho\right), \quad (21)$$

where the binary decision function $\mathcal{H}$ is approximated by a sigmoid function, allowing the gradient to be analytically calculated. Finally, Θ_E and Θ_D are updated to increase $\mathcal{J}^{\mathrm{NP}}(\Theta)$ by gradient ascent:

$$\Theta_E \leftarrow \Theta_E + \lambda \nabla_{\Theta_E} \mathcal{J}^{\mathrm{NP}}(\Theta), \quad (22)$$

$$\Theta_D \leftarrow \Theta_D + \lambda \nabla_{\Theta_D} \mathcal{J}^{\mathrm{NP}}(\Theta). \quad (23)$$

Third, to update the PDF of the latent vectors of normal sounds $p(\mathbf{z} \mid \Upsilon, y=0)$, when (18)–(23) is repeated a certain number of times, Υ is updated using the expectation-maximization (EM) algorithm for GMM using all training data of normal sounds. The above algorithm is run a pre-defined number of epochs.

D. Detailed Description of Detection Procedure

After training Θ_E and Θ_D, we can identify whether the observed sound is a normal one or not. First, the input vector $\mathbf{x}_\tau, \tau \in \{1, ..., T\}$ is calculated from the observed sound. Then, the anomaly score is calculated as (8). Finally, a decision score, $V = \frac{1}{T}\sum_{\tau=1}^{T} \mathcal{H}(\mathbf{x}_\tau, \Theta, \phi)$, is calculated, and when V exceeds a pre-defined value ϕ_V, the observed sound is determined to be anomalous. In this study, we used $\phi_V = 0$, meaning that, if the anomaly score exceeds the threshold even for one frame, the observed sound is determined to be anomalous.

E. Modified Implementation as an AUC Maximization

The receiver operating characteristic (ROC) curve and the AUC are widely used performance measures for imbalanced data classification and/or anomaly detection. The AUC is calculated as

$$\mathrm{AUC}(\Theta) = \mathbb{E}\left[\underbrace{\mathbb{E}\left[\mathcal{H}(\mathbf{x}', \Theta, \mathcal{A}(\mathbf{x}, \Theta))\right]_{\mathbf{x}'|y\neq 0}}_{\mathrm{TPR}(\Theta, \mathcal{A}(\mathbf{x}, \Theta))}\right]_{\mathbf{x}|y=0}, \quad (24)$$

$$\approx \frac{1}{M} \sum_{n=1}^{M} \mathrm{TPR}\left(\Theta, \mathcal{A}\left(\mathbf{x}_n^{(u)}, \Theta\right)\right). \quad (25)$$

As we can see in (25), anomalous sound data are needed to calculate the AUC. Although the AUC has been used as an objective function in imbalanced data classification [31]–[33], it has not been applied to unsupervised-ADS so far. Fortunately, since the proposed rejection sampling can simulate anomalous sound data, AUC maximization can be used as an objective function of ADS. Instead of $\mathcal{J}^{\mathrm{NP}}(\Theta)$, the following objective function can be used in the training procedure:

$$\mathcal{J}^{\mathrm{AUC}}(\Theta) = \frac{1}{M} \sum_{n=1}^{M} \mathrm{TPR}\left(\Theta, \mathcal{A}\left(\mathbf{x}_n^{(u)}, \Theta\right)\right) - \mathrm{FPR}\left(\Theta, \mathcal{A}\left(\mathbf{x}_n^{(u)}, \Theta\right)\right). \quad (26)$$

Hereafter, we call the proposed method using $\mathcal{J}^{\mathrm{AUC}}(\Theta)$ instead of $\mathcal{J}^{\mathrm{NP}}(\Theta)$ AUC-PROP.

IV. Experiments

We conducted experiments to evaluate the performance of the proposed method. First, we conducted an objective experiment using synthetic anomalous sounds (Section IV-B). To generate a large enough anomalous dataset for the ADS accuracy evaluation, we used collision and sustained sounds from datasets for *detection and classification of acoustic scenes and events 2016* (DCASE-2016 [36]). To show the effectiveness of the method in real environments, we conducted verification experiments in three real environments (Section IV-C).

A. Experimental Conditions

1) Compared Methods: The proposed methods described in Section III-C (NP-PROP) and Section III-E (AUC-PROP) were compared with three state-of-the-art ADS methods:

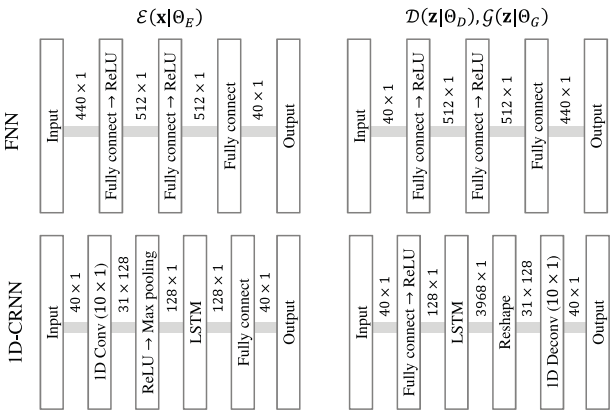


Fig. 6. Network architectures of encoder, decoder and generator used for NP-PROP, and AUC-PROP. The encoder and decoder of AE have the same architecture. In VAE, VAEGAN and CONV-PROP, the encoder has two output layers for the mean and variance vector. In VAEGAN, the architecture of the discriminator is the same as that of the encoder, but the output dimension of the fully connected layer is 1.

- AE [20]: ADS using the autoencoder described in Section II-B. The encoder and decoder were trained to minimize (9).
- VAE [24]: $\mathcal{E}$ and $\mathcal{D}$ were implemented using VAE. The encoder estimated the mean and variance parameters of the Gaussian distribution in the latent space. Then, the latent vectors were sampled from the Gaussian distribution whose parameters were estimated by the encoder. Then, the decoder reconstructed the input vector from the sampled latent vectors. Finally, the reconstruction error was calculated and used as the anomaly score.
- VAEGAN [27]: To investigate the effectiveness of the anomalous sound simulation, VAEGAN [27] was used to simulate fake normal data. The generators (i.e. VAE) were used to simulate fake normal sounds. The output of the discriminator without the sigmoid activation was used as the anomaly score.

We also used our previous work [8] (CONV-PROP) for comparison. This method uses a VAE to extract latent vectors as acoustic features. A GMM is used for the normal model, and the encoder and decoder are trained to maximize (11).

2) DNN Architecture and Setup: We tested two types of network architecture as shown in Fig. 6. The first architecture, "FNN", consisted of fully connected DNNs with three hidden layers and 512 hidden units. The rectified linear unit (ReLU) was used as the activation functions of the hidden layers. The input vector $\mathbf{x}$ was defined as

$$\mathbf{x}_\tau := \left(\ln\left[\mathrm{Mel}\left[\mathrm{Abs}\left[\mathbf{X}_{\tau-C}\right]\right]\right], ..., \ln\left[\mathrm{Mel}\left[\mathrm{Abs}\left[\mathbf{X}_{\tau+C}\right]\right]\right]\right)^\top,$$

$$\mathbf{X}_\tau := \left(X_{1,\tau}, ..., X_{\Omega,\tau}\right),$$

where $X_{\omega,\tau}$ is the discrete Fourier transform (DFT) spectrum of the observed sound, $\omega \in \{1, ..., \Omega\}$ denotes the frequency index, $C(=5)$ is the context window size, and Mel[·] and Abs[·] denote 40-dimensional Mel matrix multiplication and the element-wise absolute value. Thus, the dimension of $\mathbf{x}$ was $\mathsf{Q} = 40 \times (2C + 1) = 440$. The second architecture,"1D-CRNN", consisted in a one-dimensional convolution neural network (1D-CNN) layer and a long short-term memory (LSTM) layer; it worked well in supervised anomaly detection (race SED) in DCASE 2017 [10]. In order to detect anomalous sounds in real time, we changed the backward LSTM to a forward one. In addition, to avoid overfitting, we used only one forward LSTM layer instead of two backward LSTM layers. The input vector $\mathbf{x}$ was a 40-dimensional log mel-band energy:

$$\mathbf{x}_\tau := \ln\left(\mathrm{Mel}\left[\mathrm{Abs}\left[\mathbf{X}_\tau\right]\right]\right)^\top.$$

The dimension of $\mathbf{x}$ was $\mathsf{Q} = 40$. For each architecture, the dimension of the latent vector $\mathbf{z}$ was $\mathsf{R} = 40$. All input vectors were mean-and-variance normalized using the training data statistics.

As an implementation for the gradient method, the Adam method [34] was used instead of the gradient descent/ascent shown in (18)–(23). To avoid overfitting, L_2 normalization [35] with a regularization penalty of 10^{-4} was used. The minibatch size for all methods was $M = 512$. All models were trained for 500 epochs. In all methods, the average value of the loss was calculated on the training set at every epoch, and when the loss did not decrease for five consecutive epochs, the stepsize was decreased by half.

3) Other Conditions: All sounds were recorded at a sampling rate of 16 kHz. The frame size of the DFT was 512, and the frame was shifted every 256 samples. For $p(\mathbf{z} \mid \Upsilon, y = 0)$, the number of Gaussian mixtures was $K = 16$ and a diagonal covariance matrix was used to prevent the problem from being ill-conditioned. The EM algorithm for the GMM involved iterating (18)–(23) 30 times. All the above-mentioned conditions are summarized in Table I.

TABLE I
EXPERIMENTAL CONDITIONS

Parameters for signal processing	
Sampling rate	16.0 kHz
FFT length	512 pts
FFT shift length	256 pts
Number of mel-filterbanks	40
Other parameters	
Context window size C	5
Dimension of input vector Q for FNN	440
Dimension of input vector Q for 1D-CRNN	40
Dimension of acoustic feature vector R	40
GMM update per gradient method	30
Number of mixtures K	16
Minibatch size M	512
FPR parameter ρ	0.2
Step size λ	10^{-4}
L_2 normalization parameter	10^{-4}

B. Objective Experiments on Synthetic Data

1) Dataset: Sounds emitted from a condensing unit of an air conditioner operating in a real environment were used as the normal sounds. In addition, various machine sounds were recorded from other machines, including a *compressor*, *engine*, *compression pump*, and an *electric drill*, as well as environmental noise of factories. The normal and various machine sound data totaled 4 and 20 hours (=4 hours normal +16 hours other machines),

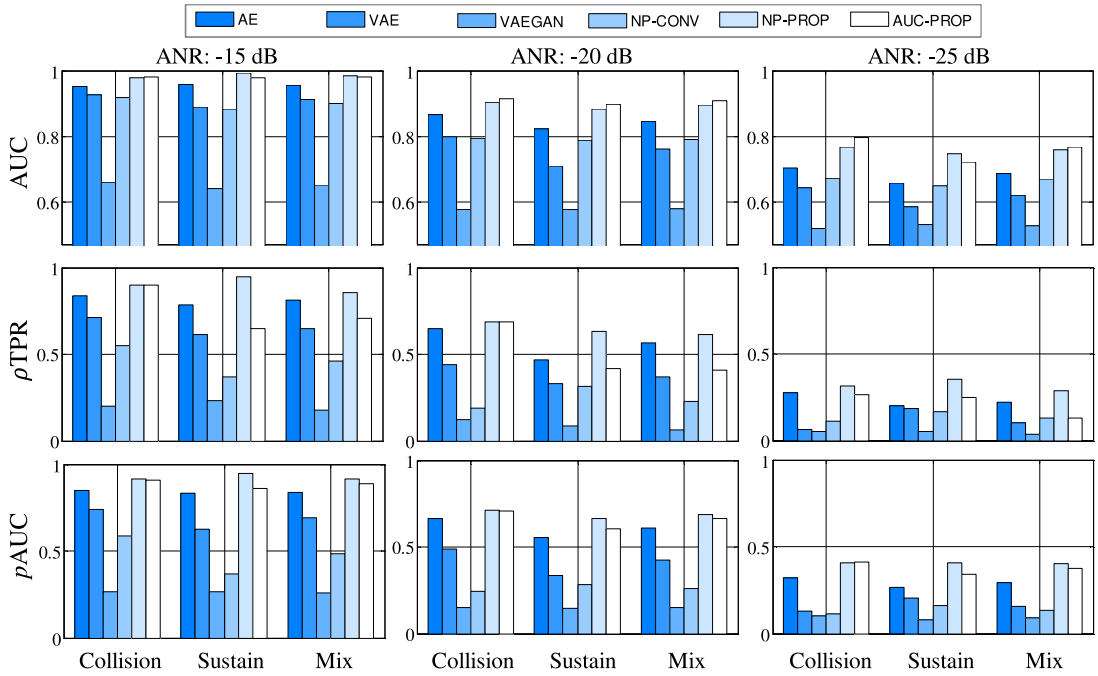


Fig. 7. Evaluation results of FNN.

respectively. These sounds were recorded at a 16-kHz sampling rate. In order to improve the robustness for different loudness levels and ratios of the normal and anomalous sound, the various machine sounds in the training dataset were augmented with a multiplication of five amplitude gains. These gains are calculated so that the maximum amplitudes of various sounds becomes to 1.0, 0.5, 0.25, 0.125, and 0.063.

Since it is difficult to collect a massive amount of test data including anomalous sounds, synthetic anomalous data were used in this evaluation. In particular, we used the training datasets for task of DCASE-2016 [36] as anomalous sounds. Although these sounds are "normal" sounds in an office, in unsupervised-ADS, the unknown sounds are categorized as "anomalous". Thus, we consider that this evaluation can at least evaluate the detection performance for unknown sounds. Since the anomalous sounds of machines are roughly categorized into collision sounds (e.g., the sound of a metal part falling on the floor) and sustained sounds (e.g., frictional sound caused by scratched bearings), we selected 80 collision sounds, including (*slamming doors* , *knocking at doors*, *keys put on a table*, *keystrokes on a keyboard*), and 60 sustained sounds (*drawers being opened*, *pages being turned*, and *phones ringing*), from this dataset [37]. To synthesize the test data, the anomalous sounds were mixed with normal sounds at anomaly-to-normal power ratios (ANRs[1]) of -15, -20 and -25 dB using the following procedure:

1) select an anomalous sound and randomly cut a normal so that has the same signal length of the selected anomalous sound.
2) for the cut normal and anomalous sounds, calculate the frame-wise log power of each of 512 points with a 256 point shift on a dB scale, namely $\mathcal{P}_\tau = 20\log_{10}\sum_{\omega=1}^{\Omega}\left|X_{\omega,\tau}\right|$.
3) select the median of $\mathcal{P}_\tau$ as the representative power of each sound as.
4) manipulate the power of the anomalous sound so that the ANR has the desired value.
5) used the cut normal sound as the test data of normal sound, and generate the test data of the anomalous sound by mixing the anomalous sound with the quarried normal sound.

In total, we used 140 normal and anomalous sound samples for each ANR condition. The training dataset of normal sounds and the MATLAB code to generate the test dataset are freely available on the website[2].

2) Results: To evaluate the performance of ADS, we used the AUC, ρ TPR, and partial AUC (p AUC) [38]. The AUC is a traditional performance measure of anomaly detection. The other two measurements evaluated the performance under low FPR conditions. ρ TPR is the TPR under the condition that FPR equals ρ. The p AUC is an AUC calculated with FPRs ranging from 0 to p with respect to the maximum value of 1. The parameters were $\rho = 0.05$ and $p = 0.1$. We evaluated these metrics for three different evaluation sets: 80 collision sounds (Collision), 60 sustained sounds (Sustain), and the sum of these $80 + 60 = 140$ sounds (Mix).

The results for each score, sound category, and ANR on FNN and 1D-CRNN are shown in Fig. 7 and Fig. 8. Overall, the

[1]ANR is a measure comparing the level of an anomalous sound to the level of a normal sound. This definition is the same as the signal-to-noise ratio (SNR) when the signal is an anomalous sound and the noise is a normal sound.

[2]https://archive.org/details/ADSdataset

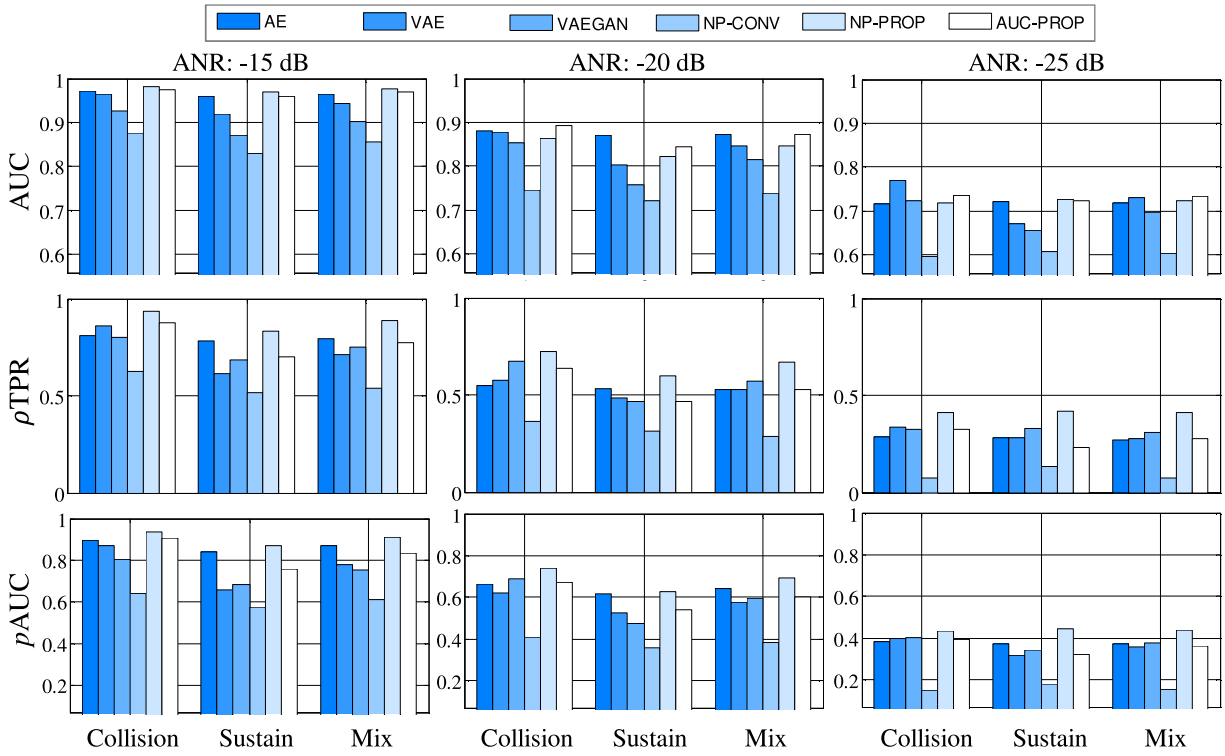


Fig. 8. Evaluation results of 1D-CRNN.

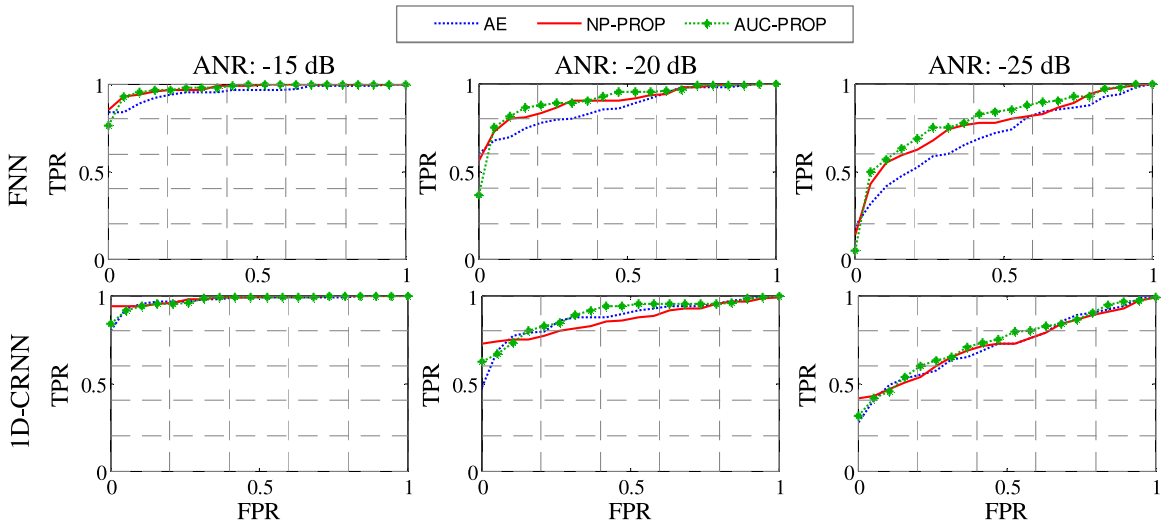


Fig. 9. ROC curves of AE, NP-PROP and AUC-PROP for each ANR condition evaluated on Mix dataset.

performances of AE, NP-PROP and AUC-PROP were better than those of VAE and VAEGAN. In detail, AE achieved high scores for all measurements, AUC-PROP achieved high scores for AUC and p AUC, and NP-PROP achieved high scores for ρ TPR and p AUC. In addition, for all conditions, the ρ TPR and p AUC scores of NP-PROP were higher than those of AE. To discuss the difference between the objective functions of AE, NP-PROP and AUC-PROP, we show the ROC curves in Fig. 9. Since the differences between the results of Collision, Sustained, and Mix were small, we plotted only those of the Mix dataset. From these ROC curves, we can see that the TPRs of NP-PROP under the low FPR conditions were significantly higher than those of other methods. This might be because the objective function of NP-PROP works to increase TPR under the low FPR condition. In addition, although AUC-PROP's TPRs under the low FPR condition were lower than those of NP-PROP, the TPRs under the moderate and high FPR conditions were higher than those of the other methods. This might be because the objective function of AUC-PROP works to increase TPR for all FPR conditions. Since the individual results and

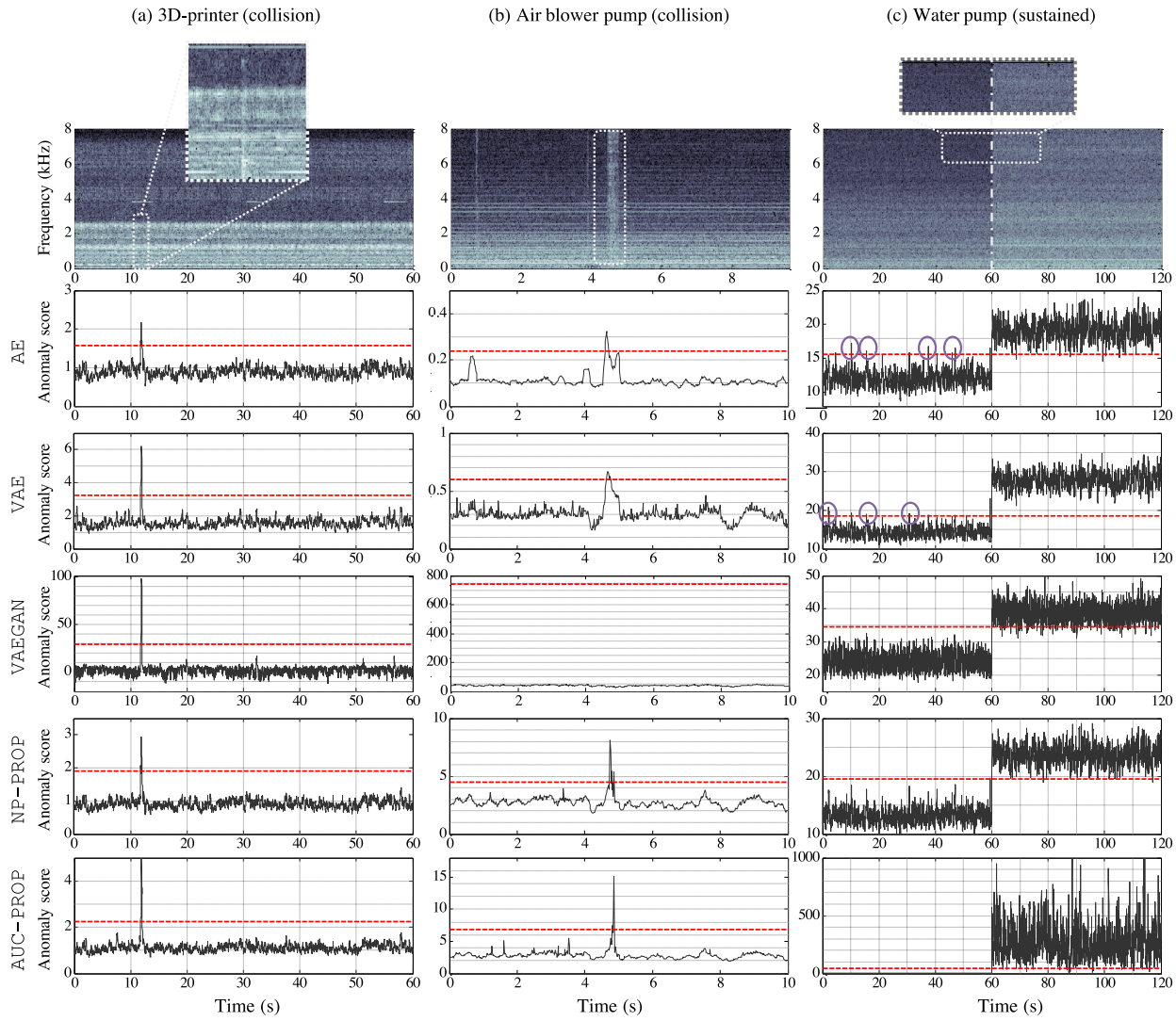


Fig. 10. Anomaly detection results for sound emitted from 3D-printer (left), air blower pump (center), and water pump (right). The top figure shows the spectrogram, and the bottom figures show the anomaly score (black solid line) and threshold $\phi_{0.001}$ (red dashed line) of each method. Anomalous sounds are enclosed in white dotted boxes, and false-positive detections are circled in purple. Since the spectrum changes due to the anomalous sounds of 3D-printer and water pump are difficult to see, their anomalous sounds are enlarged. In addition, since anomalous sound of the water pump is a sustained, 60 seconds of normal sounds and 60 seconds of anomalous sound are concatenated for comparison.

objective function tend to coincide, we consider that the training of each neural network succeeded. In addition, TPR under the low FPR conditions is especially important when the ADS is used in real environments, because if an ADS system frequently gives false alert, we cannot trust it. Therefore, unsupervised-ADS using an AE trained using (11) would be effective in real situations.

In addition, regarding the FNN results, VAE scored lower than AE, and VAEGAN scored lower than all the other methods. These results suggest that when calculating the anomaly score using a simple network architecture like FNN, a simple reconstruction error would be better than complex calculation procedures such as VAE and VAEGAN. Moreover, the scores of NP-CONV were lower than those of the DNN-based methods. In our previous study [8], we used a DNN a feature extractor and constructed the normal model by using a GMM. These results suggest that using a DNN for the normal model would be better than using a GMM.

C. *Verification Experiment in a Real Environment*

We conducted three verification experiments to test whether anomalous sounds in real environments can be detected. The target equipment and experimental conditions were as follows:

- Stereolithography 3D-printer: We collected an actual collision-type anomalous sound. Two hours worth of normal sounds were collected as training data. The anomalous sound was caused by collision of the sweeper and the formed object. The 3D-printer stopped 5 minutes after this anomalous sound occurred.
- Air blower pump: We collected an actual collision-type anomalous sound. Twenty minutes worth of normal sounds

were collected as training data. The anomalous sound was caused by blockage by a foreign object stuck in the air blower duct. This anomaly does not lead to immediate machine failure; however, it should be addressed.

- Water pump: We collected an actual sustained type anomalous sound. Three hours worth of normal sounds were collected as training data. Above 4 kHz, the anomalous sound has a larger amplitude than that of the normal sounds, and it was due to wearing of the bearings. An expert conducting a periodic inspection diagnosed that the bearings needed to be replaced.

All anomalous and normal sounds were recorded at a 16-kHz sampling rate. The other conditions were the same as in the objective experiment. The FNN architecture was used for the anomaly score calculation.

Figure 10 shows the spectrogram (top) and anomaly scores of each method (bottom). The red dashed line in each of the bottom figures is the threshold $\phi_{0.001}$, which is defined such that the FPR of the training data was 0.1%. Anomalous sounds are enclosed in white dotted boxes in the spectrograms, and the false-positive detections are circled in purple in the anomaly score graphs. Since the anomalous sound of the water pump is a sustained sound, for ease of comparison, 60 seconds of normal sounds and 60 seconds of anomalous sound are concatenated in each figure. In addition, the anomalous sounds are enlarged, since the spectrum changes due to the anomalous sounds of the 3D-printer and water pump are difficult to see.

All of the results for NP-PROP and AUC-PROP indicate that anomalous sounds were clearly detected; the anomaly scores of the anomalous sounds evidently exceeded the threshold, while those of the normal sounds were below the threshold. Meanwhile, in the results of AE and VAE, although the anomaly scores of all anomalous sounds exceeded the threshold, false-positives were also observed in the results for the water pump. In addition, although AE's anomaly score of the 3D-printer and VAE's anomaly score of the air blower pump exceeded the threshold, the excess margin of the anomaly score is small and it is difficult to use a higher threshold for reducing FPR. This problem might be because that the objective functions do not work to increase anomaly scores for anomalous sounds, and thus, the encoder and decoder reconstructed not only normal sounds but also anomalous sounds. In VAEGAN, the anomaly scores of the 3D-printer and the water pump exceeded the threshold, whereas those of the air blower pump did not exceed the threshold. The reason might be that when the generator precisely generates "fake" normal sounds, the normal model is trained to increase the anomaly scores of normal sounds. Therefore, the threshold of the air blower pump, which is defined as the FPR of normal training data becoming 0.001, takes a very high value. These verification experiments suggest that the proposed method is effective at identifying anomalous sounds under practical conditions.

V. Conclusions

This paper proposed a novel training method for unsupervised-ADS using an AE for detecting unknown anomalous sound. The contributions of this research are as follows: 1) by considering outlier-detection-based ADS as a statistical hypothesis test, we defined an objective function that builds upon the Neyman-Pearson lemma [29]. The objective function increases the TPR under a low FPR condition, which is often used in practice. 2) By considering the set of anomalous sounds to be complement to the set of normal sounds, we formulated a rejection sampling algorithm to simulate anomalous sounds. Experimental results showed that these contributions enabled us to construct an ADS system that accurately detects unknown anomalous sounds in three real environments.

In future, we will tackle the following remaining issues of ADS systems in real environments:

1) Extension to a supervised approach to detect both known and unknown anomalous sounds: while operating an ADS system in a real environment, we may occasionally obtain partial samples of anomalous sounds. While it might be better to use the collected anomalous sounds in training, the cross-entropy loss would not be the best way to detect both known and unknown anomalous sounds [39]. In addition, if we calculate the TPR in $\mathcal{J}^{\mathrm{NP}}(\Theta)$ and/or $\mathcal{J}^{\mathrm{AUC}}(\Theta)$ only using a part of the anomalous sounds, this training does not guarantee the performance for unknown anomalous sounds. Thus, we should develop a supervised-ADS method that can also detect unknown anomalous sounds; a preliminary study on this has been published in [25].

2) Incorporating machine or context-specific knowledge: to simplify the experiments, we used the simple detection rule described in Section III-D. However, for the anomaly alert, it would be better to use machine/context-specific rules, such as modifying or smoothing the detection result from the raw anomaly score. Thus, it will be necessary to develop rules or a trainable post-processing block to modify the anomaly score.

Appendix

A. List of Symbols

1. Functions

$\mathcal{J}$	Objective function.
$\mathcal{A}$	Anomaly score.
$\mathcal{H}$	Binary decision.
$\mathcal{E}$	Encoder of autoencoder.
$\mathcal{D}$	Decoder of autoencoder.
$\mathcal{G}$	Generator.
$\mathcal{N}$	Gaussian distribution.
$\mathbb{E}[\cdot]_x$	Expectation with respect to x.
$\nabla_x(\cdot)$	Gradient with respect to x.
$\mathrm{tr}(\cdot)$	Trace of matrix.
$D(A\|\|B)$	Kullback-Leibler divergence between A and B.
$\|\cdot\|_2$	L_2 norm.
$\lfloor\cdot\rfloor$	Flooring function.

2. Parameters

Θ	Parameters of normal model.
Θ_E	Parameters of encoder.
Θ_D	Parameters of decoder.
Θ_G	Parameters of generator.
Υ	Parameters of Gaussian mixture model.

3. Variables

$\mathbf{x}$	Input vector.
y	State variable.
$\mathbf{z}$	Latent vector.
ϕ	Threshold for anomaly score.
ρ	Desired false positive rate.
$\boldsymbol{\mu}$	Mean vector.
$\boldsymbol{\Sigma}$	Covariance matrix.
w	Mixing weight of Gaussian mixure model.
K	Number of gaussian mixtures.
T	Number of time frames of observation.
N	Number of training samples.
M	Minibatch size.
Q	Dimension of input vector.
R	Dimension of latent vector.
λ	Step size for gradient method.
C	Context window size.
ℓ	Temporary variable of anomaly score.
V	Anomaly decision score for one audio clip.

4. Notations

τ	Time-frame index of observation.
n	Index of training sample.
k	Index of Gaussian distribution.
$(\cdot)^\top$	Transpose of matrix or vector.
$(\cdot)^{(u)}$	Variable of normal sound.
$(\cdot)^{(a)}$	Variable of anomalous sound.
$(\cdot)^{(v)}$	Variable of various sound.

REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1306–1309.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2007, pp. 21–26.

[3] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 713–719, Aug. 2011.

[4] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, Jan. 2016.

[5] P. Coucke, B. De Ketelaere, and J. De Baerdemaeker, "Experimental analysis of the dynamic, mechanical behavior of a chicken egg," *J. Sound Vib.*, vol. 266, pp. 711–721, 2003.

[6] Y. Chung, S. Oh, J. Lee, D. Park, H. H. Chang, and S. Kim, "Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems," *Sensors*, vol. 13, pp. 12929–12942, 2013.

[7] A. Yamashita, T. Hara, and T. Kaneko, "Inspection of visible and invisible features of objects with image and sound signal processing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 3837–3842.

[8] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman–Pearson lemma," in *Proc. 25th Eur. Signal Process. Conf.*, 2017, pp. 698–702.

[9] A. Mesaros *et al.*"DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. Detection Classification Acoustic Scenes Events 2017 Workshop (DCASE2017)*, 2017, pp. 85–92.

[10] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2017.

[11] E. Cakir and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2017.

[12] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural I-vectors and deep convolutional neural networks," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2016.

[13] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using Svm hyperplane," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2017.

[14] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2016.

[15] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2017.

[16] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification and domestic audio tagging," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2016.

[17] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, pp. 85–126, 2004.

[18] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *J. Comput. Netw.*, vol. 51, pp. 3448–3470, 2007.

[19] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, 2009, Art. no. 15.

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

[21] T. Tagawa, Y. Tadokoro, and T. Yairi, "Structured denoising autoencoder for fault detection and analysis," in *Proc. 6th Asian Conf. Mach. Learn.*, 2015, pp. 96–111.

[22] E. Marchi, F. Vesperini, F. Weninger, F. Eyben, S. Squartini, and B. Schuller, "Non-linear prediction with LSTM recurrent neural networks for acoustic novelty detection," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–7.

[23] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[24] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," SNU Data Mining Center, Seoul, South Korea, Tech. Rep., 2015, pp. 1–18.

[25] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2336–2370.

[26] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[27] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1558–1566.

[28] T. Schlegl, P. Seebock, S. M. Waldstein, U. S. Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2017, pp. 146–157.

[29] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philos. Trans. Roy. Soc. London*, vol. 231, pp. 289–337, 1933.

[30] G. Casella and R. L. Berger, "Section 8.3.2: Most powerful tests," in *Statistical Inference*. Pacific Grove, CA, USA: Duxbury, 2001, pp. 387–393.

[31] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[32] A. Herschtal and B. Raskutti, "Optimising area under the ROC curve using gradient descent," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 49.

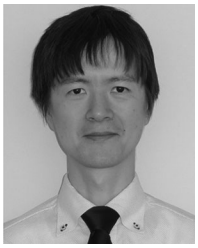
[33] A. Fujino and N. Ueda, "A semi-supervised AUC optimization method with generative models," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 883–888, p. 49.

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

[35] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. 4th Int. Conf. Neural Inf. Process. Syst.*, 1991, pp. 950–957.

[36] "DCASE 2016," http://www.cs.tut.fi/sgn/arg/dcase2016/, Accessed on. Oct. 29, 2018.
[37] "Download - DCASE 2016," http://www.cs.tut.fi/sgn/arg/dcase2016/, Accessed on. Oct. 29, 2018.
[38] S. D. Walter, "The partial area under the summary ROC curve," *Statist. Medicine*, vol. 24, pp. 2025–2040, 2005.
[39] N. Gornitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *J. Artif. Intell. Res.*, vol. 46, pp. 235–262, 2013.

Yuma Koizumi (M'15) received the B.S. and M.S. degrees in computer and information sciences from Hosei University, Tokyo, Japan, in 2012 and 2014, and the Ph.D. degree in computer and information sciences from the University of Electro-Communications, Tokyo, Japan, in 2017. Since 2014, he has been with Nippon Telegraph and Telephone Corporation (NTT), where he has been involved with research on acoustic signal processing and machine learning, including basic research of sound source enhancement and unsupervised/supervised ADSs. Dr. Koizumi was the recipient of the FUNAI Best Paper Award and the Information Processing Society of Japan (IPSJ) Yamashita SIG Research Award from the IPSJ in 2013 and 2014, respectively, and the Awaya Prize from the Acoustical Society of Japan (ASJ) in 2017. He is a member of the ASJ and the Institute of Electronics, Information and Communication Engineers.



Shoichiro Saito (S'06–M'07) received the B.E. and M.E. degrees in information science from the University of Tokyo, Tokyo, Japan, in 2005 and 2007. Since joining NTT Corporation, Tokyo, Japan, in 2007, he has been involved with the research and development of acoustic signal processing systems, including acoustic echo cancellers, hands-free telecommunication, and ADS. He is currently a Senior Research Engineer with the Audio, Speech, and Language Media Laboratory, NTT Media Intelligence Laboratories, NTT Corporation. Mr. Saito is a member of the Institute of Electronics, Information and Communication Engineers and the Acoustical Society of Japan.

Hisashi Uematsu received the B.E., M.E., and Ph.D. degrees in information science from Tohoku University, Sendai, Japan, in 1991, 1993, and 1996, respectively. He joined NTT Corporation, Tokyo, Japan, in 1996 and has been engaged in research on psychoacoustics (human auditory mechanisms) and digital signal processing. He is currently a Senior Research Engineer with the Cross-Modal Computing Project, NTT Media Intelligence Laboratories, NTT Corporation. Dr. Uematsu was the recipient of the Awaya Prize from the Acoustical Society of Japan (ASJ) in 2001. He is a member of the ASJ.

Yuta Kawachi received the B.E. and M.E. degrees in information science from Waseda University, Tokyo, Japan, in 2012 and 2014, respectively. Since joining NTT Corporation, Tokyo, Japan, in 2014, he has been involved with research on acoustic signal processing and machine learning. Mr. Kawachi is a member of the Acoustical Society of Japan.

Noboru Harada (M'99–SM'18) received the B.S. and M.S. degrees in computer science from the Department of Computer Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, Japan, in 1995 and 1997, respectively, and the Ph.D. degree in computer science from the Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan, in 2017. Since joining NTT Corporation, Tokyo, Japan, in 1997, he has been involved with research on speech and audio signal processing, such as high efficiency coding and lossless compression. His current research interests include acoustic signal processing and machine learning for acoustic event detection, including ADS. Dr. Harada was the recipient of the Technical Development Award from the Acoustical Society of Japan (ASJ) in 2016, the Industrial Standardization Encouragement Awards from the Ministry of Economy Trade and Industry of Japan in 2011, and the Telecom System Technology Paper Encouragement Award from the Telecommunications Advancement Foundation of Japan in 2007. He is a member of the ASJ, the Institute of Electronics, Information and Communication Engineers, and the Information Processing Society of Japan.