Optimizing Acoustic Feature Extractor for Anomalous Sound Detection Based on Neyman-Pearson Lemma

Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, and Noboru Harada
NTT Media Intelligence Laboratories, NTT Corporation, Tokyo, Japan

***Abstract*—We propose a method for optimizing an acoustic feature extractor for anomalous sound detection (ASD). Most ASD systems adopt outlier-detection techniques because it is difficult to collect a massive amount of anomalous sound data. To improve the performance of such outlier-detection-based ASD, it is essential to extract a set of efficient acoustic features that is suitable for identifying anomalous sounds. However, the ideal property of a set of acoustic features that maximizes ASD performance has not been clarified. By considering outlier-detection-based ASD as a statistical hypothesis test, we defined optimality as an objective function that adopts Neyman-Pearson lemma; the acoustic feature extractor is optimized to extract a set of acoustic features which maximize the true positive rate under an arbitrary false positive rate. The variational auto-encoder is applied as an acoustic feature extractor and optimized to maximize the objective function. We confirmed that the proposed method improved the F-measure score from 0.02 to 0.06 points compared to those of conventional methods, and ASD results of a stereolithography 3D-printer in a real-environment show that the proposed method is effective in identifying anomalous sounds.**

***Index Terms*—Anomalous sound detection, acoustic feature, objective function, deep neural network, Gaussian mixture model.**

I. Introduction

Much attention has recently been on anomalous sound detection (ASD) such as audio surveillance [1], [2], [3], [4] and equipment inspection [5], [6]. The goal with ASD is to prevent accidents and/or mechanical failures by detecting sounds that do not normally occur, *i.e.,* anomalous sound [7]. In this study, we investigated an ASD for industrial equipment by focusing on machine-operating sounds.

Since anomalous sound due to equipment failure rarely occurs, it is difficult to collect a massive amount of anomalous-sound data [7]. Therefore, most anomaly-detection systems adopt an outlier-detection technique (Fig. 1) [8], [9], [10], [11]. In outlier-detection-based ASD, the deviance between the normal model and a set of acoustic features extracted from an observed sound is calculated (*i.e., anomaly score*). The observed sound is identified as an anomalous sound when the anomaly score is higher than the pre-defined threshold value. Therefore, it is essential to extract a set of *informative* acoustic features, which provides a small anomaly score for normal sound and large anomaly score for anomalous sound.

To extract a set of acoustic features for the sound-identification problem, feature-extractor-optimization methods have been actively investigated [12], [13], [14]. These studies have revealed that it is necessary to determine both spectral and temporal characteristics to accurately identify various sounds. In recent years, deep neural networks (DNNs) have been used as a feature extractor to determine spectro-temporal characteristics [15], [16], and Chakrabarty *et al.* have reported that the restricted-Boltzmann-machine-based feature extractor is efficient for improving ASD accuracy [17]. Although determining the spectro-temporal characteristics is a requirement of acoustic features, the optimal acoustic features to maximize ASD performance has not been clarified.

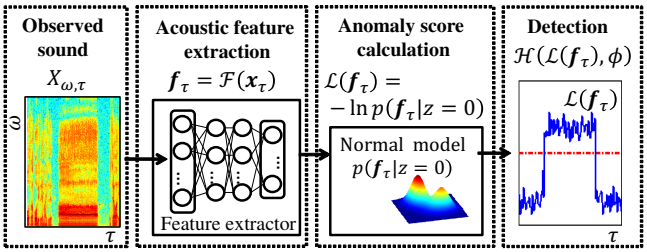

Fig. 1. Procedure of anomaly sound detection based on outlier detection

We propose a method for optimizing an acoustic feature extractor for anomalous sound detection. By considering outlier-detection-based ASD as a statistical hypothesis test, we defined optimality as an objective function that adopts Neyman-Pearson lemma [18]; the acoustic feature extractor is optimized to extract a set of acoustic features which maximize the true positive rate under an arbitrary false positive rate. A DNN-based feature extractor is optimized with the proposed method by applying the variational auto-encoder (VAE) [19]. We experimentally show that ASD performance improves by optimizing the DNN-based feature extractor using the proposed method.

The rest of this paper is organized as follows. Section II briefly introduces outlier-detection-based ASD. Then, in Section III, we discuss our proposed method for optimizing a feature extractor and its implementation. After investigating the performance of the proposed method in Section IV, we conclude this paper in Section V.

II. Detection procedure of anomalous sound based on outlier-detection

Anomaly sound detection is an identification problem on whether the sound emitted from the target machine $X_{\omega,\tau} \in \mathbb{C}^{\Omega\times T}$ is a normal sound or anomalous one. Here $\omega = \{1,2,...,\Omega\}$ and $\tau = \{1,2,...,T\}$ denote the frequency and time indices, respectively. In this section, we briefly introduce the procedure of outlier-detection-based ASD (Fig. 1).

ISBN 978-0-9928626-7-1 © EURASIP 2017

First, a set of acoustic features $\boldsymbol{f}_\tau \in \mathbb{R}^{\mathsf{D}}$ is extracted as

$$\boldsymbol{f}_\tau = \mathcal{F}(\boldsymbol{x}_\tau), \tag{1}$$

where $\mathcal{F}$ is an acoustic feature extractor. To determine the spectro-temporal characteristics of the observed sound, the input vector $\boldsymbol{x}_\tau$ is obtained by concatenating several frames of observation by accounting for previous and future frames, as $\boldsymbol{x}_\tau = \left(\boldsymbol{X}_{\tau-P_b}, \boldsymbol{X}_{\tau-P_b+1}, ..., \boldsymbol{X}_{\tau+P_f}\right)^\top$, where $\boldsymbol{X}_\tau = \ln\left(|X_{1,\tau}|, |X_{2,\tau}|, ..., |X_{\Omega,\tau}|\right)$, $\top$ denotes transposition, and P_b and P_f are the context window size of previous and future frames, respectively. As a simple implementation of $\mathcal{F}$, fully-connected DNN [20] can be used as

$$\mathcal{F}(\boldsymbol{x}_\tau) = \mathbf{W}^{(L)}\boldsymbol{h}_\tau^{(L-1)} + \mathbf{b}^{(L)}, \tag{2}$$

$$\boldsymbol{h}_\tau^{(l)} = \sigma_\theta\left\{\mathbf{W}^{(l)}\boldsymbol{h}_\tau^{(l-1)} + \mathbf{b}^{(l)}\right\}, \tag{3}$$

where l, L, $\mathbf{W}^{(l)}$, and $\mathbf{b}^{(l)}$ are the layer index, the number of layers, the weight matrix, and bias vector, respectively. The function σ_θ is a nonlinear activation function, such as the sigmoid function. The input vector $\boldsymbol{x}_\tau$ is passed to the first layer of the network as $\boldsymbol{h}_\tau^{(1)} = \boldsymbol{x}_\tau$.

Next, the anomaly score $\mathcal{L}(\boldsymbol{f}_\tau)$ is calculated as the negative-log-likelihood of the normal model

$$\mathcal{L}(\boldsymbol{f}_\tau) = -\ln p(\boldsymbol{f}_\tau|z=0), \tag{4}$$

where $p(\boldsymbol{f}|z=0)$ is the normal model, which is the probability density function (PDF) of the set of acoustic features extracted from normal sound, $z \in \{0, 1, ..., \infty\}$ is the index of types of machine which emitted $X_{\omega,\tau}$, and $z=0$ denotes $X_{\omega,\tau}$ is emitted from the target machine. As a simple implementation of $p(\boldsymbol{f}|z=0)$, a Gaussian mixture model (GMM) can be used as $p(\boldsymbol{f}|z=0) = \sum_{c=1}^{\mathsf{C}} w_c \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, where C is the number of mixtures, $\mathcal{N}$ is the Gaussian distribution, and w_c, $\boldsymbol{\mu}_c$, and $\boldsymbol{\Sigma}_c$ are the mixing weight, mean vector, and covariance matrix of the c-th Gaussian distribution, respectively. Finally, when $\mathcal{L}(\boldsymbol{f}_\tau)$ exceeds the pre-defined threshold value ϕ, $X_{\omega,\tau}$ is determined as anomalous sound;

$$\mathcal{H}(\mathcal{L}(\boldsymbol{f}_\tau), \phi) = \begin{cases} 0 \text{ (Normal sound)} & \mathcal{L}(\boldsymbol{f}_\tau) \leq \phi \\ 1 \text{ (Anomalous sound)} & \mathcal{L}(\boldsymbol{f}_\tau) > \phi \end{cases}. \tag{5}$$

In outlier-detection-based ASD, it is necessary to optimize $\mathcal{F}$ to extract a set of informative acoustic features, which provides a small anomaly score for normal sound and a large anomaly score for anomalous sound. However, since the optimal set of acoustic features to maximize ASD accuracy has not been clarified, the objective function for $\mathcal{F}$ has not been established.

III. Proposed method

A. Basic property of optimal set of acoustic features for anomaly sound detection

To optimize a system, engineers define the optimality of the function output, *i.e.,* objective function. The system can be optimized by feeding back the evaluation of its output in accordance with the optimality. Therefore, to optimize the system more efficiently, it is necessary to define a suitable optimality of the problem. In this section, to optimize $\mathcal{F}$, we discusses the optimality of the set of acoustic features (*i.e.,* DNN outputs) for outlier-detection-based ASD.

From (4) and (5), the observed sound is identified as an anomalous sound when the following inequality is satisfied.

$$p(\mathcal{F}(\boldsymbol{x}_\tau)|z=0) < \exp(-\phi). \tag{6}$$

Since ϕ is assumed to be sufficiently large, anomalous sound can be defined as "sound whose acoustic features cannot be regarded as a sample generated from the normal model." Then, it can be regarded as outlier-detection-based ASD, which is a statistical hypothesis test. The observed sound is identified as an anomalous sound when the following hypothesis is rejected.

Null hypotheses: the set of acoustic features $\mathcal{F}(\boldsymbol{x})$ is a sample from normal model $p(\mathcal{F}(\boldsymbol{x})|z=0)$.

Thus, we consider that the optimal property of the statistical hypothesis test can be applied to the objective function.

The Neyman-Pearson lemma [18] denotes the criterion of the most powerful hypothesis test between two simple hypotheses; the most powerful test function maximizes the true positive rate (TPR) with a constraint under the false positive rate (FPR) equals ρ. The TPR and FPR can be calculated as

$$\text{TPR}(\mathcal{F}, \phi) = \mathbb{E}\left[\mathcal{H}(\mathcal{L}(\mathcal{F}(\boldsymbol{x})), \phi)\right]_{\boldsymbol{x}|z\neq 0}, \tag{7}$$

$$\text{FPR}(\mathcal{F}, \phi) = \mathbb{E}\left[\mathcal{H}(\mathcal{L}(\mathcal{F}(\boldsymbol{x})), \phi)\right]_{\boldsymbol{x}|z=0}, \tag{8}$$

respectively, and $\mathbb{E}[\cdot]_{\boldsymbol{x}}$ is the expectation operator for $\boldsymbol{x}$. We define a threshold value ϕ_ρ that satisfies $\text{FPR}(\mathcal{F}, \phi_\rho) = \rho$, then the most powerful test function maximizes the following equation

$$\mathcal{J} = \text{TPR}(\mathcal{F}, \phi_\rho) + \{\rho - \text{FPR}(\mathcal{F}, \phi_\rho)\}. \tag{9}$$

To derivate the objective function for $\mathcal{F}$, we aim to maximize (9) with respect to $\mathcal{F}$. To simplify the problem, we consider ϕ_ρ as a constant value that is irrelevant with $\mathcal{F}$. Then, the objective function for $\mathcal{F}$ can be written as

$$\mathcal{F} \leftarrow \arg\max_{\mathcal{F}} \text{TPR}(\mathcal{F}, \phi_\rho) - \text{FPR}(\mathcal{F}, \phi_\rho). \tag{10}$$

In the following sections, a DNN-based feature extractor is optimized using (10) by applying the VAE [19].

B. Acoustic feature-extractor optimization using variational auto-encoder

To numerically optimize $\mathcal{F}$, (10) is reformed to differentiable form with respect to $\mathcal{F}$. Then, a gradient method can be used to optimize $\mathcal{F}$. First, we assume that $\mathcal{F}$ and $p(\mathcal{F}(\boldsymbol{x}_\tau)|z=0)$ are differentiable composite functions with respect to the parameters of $\mathcal{F}$, *e.g.*, full-connected DNN and GMM, respectively. Next, $\mathcal{H}(\mathcal{L}(\boldsymbol{x}_\tau), \phi)$ is approximated to differentiable form using the sigmoid function as

$$\tilde{\mathcal{H}}(\mathcal{L}(\mathcal{F}(\boldsymbol{x}))), \phi) = \frac{1}{1+\exp\{\mathcal{L}(\mathcal{F}(\boldsymbol{x})) - \phi\}}. \tag{11}$$

Then, (10) can be reformed to differentiable form with respect to the parameters of $\mathcal{F}$ as

ISBN 978-0-9928626-7-1 © EURASIP 2017

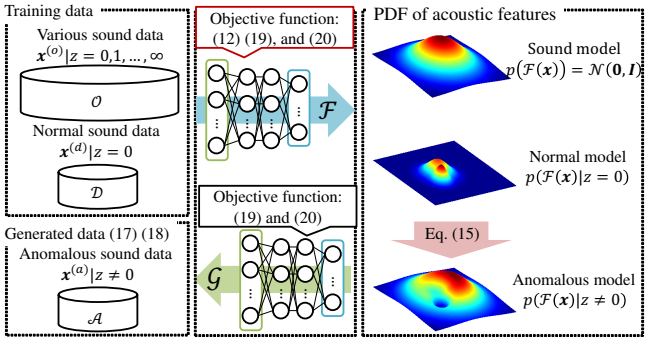

Fig. 2. Concept of optimization procedure of proposed method

$$\mathcal{F} \leftarrow \arg\max_{\mathcal{F}} \frac{1}{K_a}\sum_{k=1}^{K_a}\tilde{\mathcal{H}}(\mathcal{L}(\mathcal{F}(\boldsymbol{x}_k^{(a)})),\phi_\rho) - \frac{1}{K_d}\sum_{k=1}^{K_d}\tilde{\mathcal{H}}(\mathcal{L}(\mathcal{F}(\boldsymbol{x}_k^{(d)})),\phi_\rho), \tag{12}$$

where, to optimize $\mathcal{F}$ using training data, the expectations in the TPR and FPR are replaced with the arithmetic mean of the training data of normal sound $\boldsymbol{x}_k^{(d)}$ and anomalous sound $\boldsymbol{x}_k^{(a)}$ as

$$\mathcal{D} = \{\boldsymbol{x}_k^{(d)} \in \mathbb{R}^{\mathsf{Q}} | k = 1, ..., K_d\}, \tag{13}$$

$$\mathcal{A} = \{\boldsymbol{x}_k^{(a)} \in \mathbb{R}^{\mathsf{Q}} | k = 1, ..., K_a\}, \tag{14}$$

where $\mathsf{Q} = \Omega \times (P_b + P_f + 1)$ and K_d, K_a are the number of training samples of normal sound and anomalous sound, respectively. In (12), to satisfy $\mathrm{FPR}(\mathcal{F}, \phi_\rho) = \rho$, ϕ_ρ is set as the $\lfloor \rho K_d \rfloor$-th value of sorted $\mathcal{L}(\mathcal{F}(\boldsymbol{x}^{(d)}_{1,...,K_d}))$ in descending order, where $\lfloor \cdot \rfloor$ is the floor function.

Unfortunately, since it is difficult to collect anomalous sound data, $\mathcal{A}$ would not be massive enough to approximate the expectation by arithmetic mean. To accurately calculate the arithmetic mean, anomalous sound data are generated using a sampling algorithm. In outlier detection, anomalous sound is defined as sound whose acoustic feature cannot be regarded as a sample generated from the normal model. Thus, we define the PDF of the set of acoustic features of anomalous sound $p(\mathcal{F}(\boldsymbol{x})|z \neq 0)$ as

$$\begin{aligned} p(\mathcal{F}(\boldsymbol{x})|z \neq 0) &= \sum_{i=0}^{\infty} p(\mathcal{F}(\boldsymbol{x})|z=i) - p(\mathcal{F}(\boldsymbol{x})|z=0), \\ &\approx p(\mathcal{F}(\boldsymbol{x})) - p(\mathcal{F}(\boldsymbol{x})|z=0), \end{aligned} \tag{15}$$

where the priori probability $p(z)$ was omitted. In the first term of (15), the machine type index z is marginalized. Thus, $p(\mathcal{F}(\boldsymbol{x}))$ would be regarded as the PDF of the set of acoustic features extracted from various machine sounds emitted from many other equipments recorded in other factories as

$$\mathcal{O} = \{\boldsymbol{x}_k^{(o)} \in \mathbb{R}^{\mathsf{Q}} | k = 1, ..., K_o\}. \tag{16}$$

Hence, by calculating $p(\mathcal{F}(\boldsymbol{x}))$ using $\mathcal{O}$, $p(\mathcal{F}(\boldsymbol{x})|z \neq 0)$ can be approximately calculated using (15).

In this study, $\mathcal{A}$ was generated using $p(\mathcal{F}(\boldsymbol{x})|z \neq 0)$ and an inverse function of the feature extractor $\mathcal{G}$. First, by using a sampling algorithm, the set of acoustic features of anomalous sound $\tilde{\boldsymbol{f}}_k^{(a)}$ is generated as

$$\tilde{\boldsymbol{f}}_k^{(a)} \sim p(\mathcal{F}(\boldsymbol{x})) - p(\mathcal{F}(\boldsymbol{x})|z=0), \tag{17}$$

where $\sim$ denotes sampling from the right-hand-side distribution. Next, $\mathcal{A}$ is generated using an inverse function of feature extractor $\mathcal{G}$ as

$$\boldsymbol{x}_k^{(a)} \leftarrow \mathcal{G}(\tilde{\boldsymbol{f}}_k^{(a)}). \tag{18}$$

To generate $\tilde{\boldsymbol{f}}_k^{(a)}$ and $\mathcal{A}$ easily and accurately, $\mathcal{F}$ and $\mathcal{G}$ are implemented using the VAE [19], as shown in Fig. 2. In our implementation, $\mathcal{F}(\boldsymbol{x}_k)$ outputs $\boldsymbol{f}_k$ (*i.e.* mean vector) and its variance $\boldsymbol{\sigma}(\boldsymbol{x}_k) = (\sigma_{k,1}, ..., \sigma_{k,\mathsf{D}})^\top$, and $\mathcal{F}$ and $\mathcal{G}$ are trained to minimize the following reconstruction error

$$\boldsymbol{E} = \sum_{k_o=1}^{K_o} ||\mathcal{G}(\boldsymbol{\zeta}_k^{(o)}) - \boldsymbol{x}_k^{(o)}||^2, \tag{19}$$

with a constraint as $p(\boldsymbol{f}^{(o)}|\boldsymbol{x}^{(o)}) = \mathcal{N}(\boldsymbol{f}^{(o)}|\boldsymbol{0}_\mathsf{D}, \boldsymbol{I}_\mathsf{D})$, where $\boldsymbol{0}_\mathsf{D}$ and $\boldsymbol{I}_\mathsf{D}$ are the zero-vector and identity matrix of size D, $\boldsymbol{\zeta}_k^{(o)} = \boldsymbol{f}_k^{(o)} + \boldsymbol{\sigma}(\boldsymbol{x}_k^{(o)}) \odot \boldsymbol{\epsilon}_k$, $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\boldsymbol{\epsilon}_k|\boldsymbol{0}_\mathsf{D}, \boldsymbol{I}_\mathsf{D})$, and $\odot$ denotes the element-wise product. The constraint is achieved by minimizing the Kullback-Leibler divergence (KLD) between $p(\mathcal{F}(\boldsymbol{x}))$ and $\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}_\mathsf{D}, \boldsymbol{I}_\mathsf{D})$ as

$$\begin{aligned} &\sum_{k=1}^{K_o} KL\left[p(\boldsymbol{f}_k^{(o)}|\boldsymbol{x}_k^{(o)})||\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}_\mathsf{D}, \boldsymbol{I}_\mathsf{D})\right] \\ &= \frac{1}{2}\sum_{k=1}^{K_o}\sum_{d=1}^{D}\left(1 + \ln((\sigma_{k,d}^{(o)})^2) - (f_{k,d}^{(o)})^2 - (\sigma_{k,d}^{(o)})^2\right) \end{aligned}. \tag{20}$$

In this study, instead of using (17), the simple generation algorithm shown in **Algorithm 2** was used. In addition, $\mathcal{F}$ and $\mathcal{G}$ were implemented using fully-connected DNNs, and the symmetric network architecture of $\mathcal{F}$ was used for that of $\mathcal{G}$. Then, $\mathcal{F}$ and $\mathcal{G}$ were trained to maximize (12) and to minimize (19) and (20), alternately.

Algorithm 1 Training algorithm of proposed method

Input: $\mathcal{D}$ and $\mathcal{O}$
Output: $\mathcal{F}$ and $p(\mathcal{F}(\boldsymbol{x})|z=0)$
Initialize $\mathcal{F}$, $\mathcal{G}$, and $p(\mathcal{F}(\boldsymbol{x})|z=0)$
while *repeat for designated times* **do**
  $\boldsymbol{x}_k^{(d)}$ and $\boldsymbol{x}_k^{(o)} \leftarrow$ Random draw from $\mathcal{D}$ and $\mathcal{O}$
  $\mathcal{F}$ and $\mathcal{G} \leftarrow$ Minimize (19) and (20) using $\boldsymbol{x}_k^{(o)}$
  $\boldsymbol{f}^{(d)}_{1,...,K_d} \leftarrow \mathcal{F}(\boldsymbol{x}^{(d)}_{1,...,K_d})$
  $p(\mathcal{F}(\boldsymbol{x})|z=0) \leftarrow$ GMM-EM-algo. using $\boldsymbol{f}^{(d)}_{1,...,K_d}$
  $\phi_\rho \leftarrow \lfloor \rho K_d \rfloor^{\text{th}}$ value of descend sorted $\mathcal{L}(\boldsymbol{f}^{(d)}_{1,...,K_d})$
  $\boldsymbol{x}_k^{(a)} \leftarrow$ Generate K samples using **Algorithm 2** and (18)
  $\mathcal{F} \leftarrow$ Maximize (12) $\boldsymbol{x}_k^{(d)}$ and $\boldsymbol{x}_k^{(a)}$
end while

C. Training procedure

We now describe the details of the training procedure shown in **Algorithm 1**. The algorithm inputs are training data of normal sound $\mathcal{D}$ and various sounds $\mathcal{O}$, and outputs are $\mathcal{F}$

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE I
EVALUATION RESULTS

SNR (dB)	-10 dB			-5 dB			0 dB			5 dB		
score	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
AE	0.62	0.98	0.76	**0.94**	0.76	0.84	0.95	0.87	0.91	**0.98**	0.91	0.94
VAE	0.51	**0.94**	0.67	0.52	**1.0**	0.68	0.61	**0.98**	0.75	0.86	0.84	0.85
PROP	**0.76**	0.91	**0.82**	0.84	0.96	**0.90**	**0.96**	0.89	**0.93**	0.92	**1.0**	**0.96**

Algorithm 2 Generation algorithm of anomalous sound

Input: $p(\mathcal{F}(\boldsymbol{x})|z=0)$ and ϕ_ρ
Output: $\tilde{\boldsymbol{f}}_k^{(a)}$
while $\mathcal{L}(\tilde{\boldsymbol{f}}_k^{(a)}) \leq \phi_\rho$ **do**
$\tilde{\boldsymbol{f}}_k^{(a)} \sim \mathcal{N}(\boldsymbol{0}_{\mathsf{D}}, \boldsymbol{I}_{\mathsf{D}})$
end while

and $p(\mathcal{F}(\boldsymbol{x})|z=0)$. In this study, $p(\mathcal{F}(\boldsymbol{x})|z=0)$ was implemented using a GMM.

First, K-samples of normal sounds $\boldsymbol{x}_k^{(d)}$ and various sounds $\boldsymbol{x}_k^{(o)}$ are randomly drawn from $\mathcal{D}$ and $\mathcal{O}$, *i.e.*, mini-batches. Then, $\mathcal{F}$ and $\mathcal{G}$ are updated one step to decrease (19) and (20) using $\boldsymbol{x}_k^{(o)}$ by stochastic gradient descent. Next, sets of acoustic features $\boldsymbol{f}_{1,\ldots,K_d}^{(d)}$ is extracted from all training data of normal sound $\mathcal{D}$. Then, the normal model $p(\mathcal{F}(\boldsymbol{x})|z=0)$ is updated with the expectation-maximization (EM)-algorithm for GMM using $\boldsymbol{f}_{1,\ldots,K_d}^{(d)}$. After updating the normal model, in order to set the threshold value ϕ_ρ, anomaly scores of the training data of normal sound are calculated as $\mathcal{L}(\boldsymbol{f}_{1,\ldots,K_d}^{(d)})$ and sorted in descending order. Then ϕ_ρ is set as the $\lfloor \rho K_d \rfloor$-th value of sorted anomaly scores. Finally, K-samples of anomalous sounds $\boldsymbol{x}_k^{(a)}$ are generated with **Algorithm 2** and (18), then $\mathcal{F}$ is updated one step to increase (12) using $\boldsymbol{x}_k^{(d)}$ and $\boldsymbol{x}_k^{(a)}$ by stochastic gradient ascent.

IV. EXPERIMENTS

A. Experimental conditions

We conducted experiments to evaluate the performance of the proposed method (PROP). As comparison methods, we applied the auto-encoder (AE) and VAE (VAE) for $\mathcal{F}$.

The dimension of the number of output-units of $\mathcal{F}$ was $\mathsf{D}=32$ and the context window sizes were $P_b = P_f = 10$. To avoid over-fitting, $X_{\omega,\tau}$ was compressed using $\mathsf{B}=64$ mel-filterbanks. Thus, the dimension of input $\boldsymbol{x}$ was $\mathsf{Q} = 64 \times (P_b + P_f + 1) = 1344$. The architecture of $\mathcal{F}$ was as follows: the number of hidden layers was 3, the number of units in each hidden layer was 512, and the rectified linear unit was used as the activation function. The $\mathcal{F}$ and $\mathcal{G}$ were initialized with values that follow a normal distribution. The Adam method [21] was used as a gradient method, and L_2 normalization with parameter $\lambda = 10^{-5}$ was used for weight normalization [22]. The dropout method was used with the dropout probabilities of the input and hidden layers, which were 0.2 and 0.5, respectively. The mini-batch size was $K = 100$. After 500-epoch training, the training algorithm was terminated. The FPR parameter heuristically determined

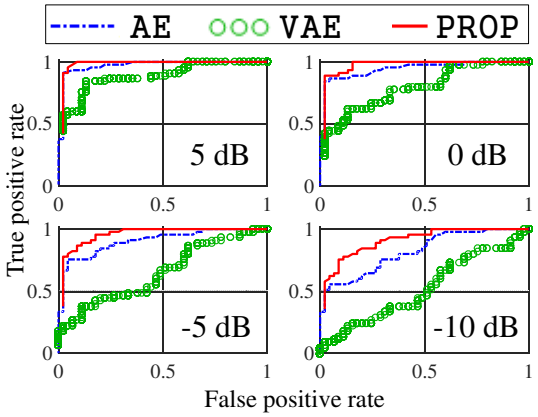

Fig. 3. ROC curves of each SNR condition

$\rho = 0.05$. For the normal model, the number of Gaussian mixtures was $\mathsf{C}=16$ and the diagonal covariance matrix was used to avoid ill-condition. Both $\mathcal{D}$ and $\mathcal{O}$ were used to train the $\mathcal{F}$ of the comparison methods.

B. Experiment data

Since it is difficult to collect a massive amount of test data including anomalous sounds, synthetic anomalous data were used for this evaluation. Normal sounds emitted from an engine in real-environment were used as training data of normal sound $\mathcal{D}$. Other-type-machine sounds recorded in other factories and machines were used as the various-sound data $\mathcal{O}$. The size of $\mathcal{D}$ and $\mathcal{O}$ were 1 and 20 hours, respectively. These sounds were recorded at a 16-kHz sampling rate. Anomalous sounds consisted of 45 machine-operating sounds; 15 sustainable sounds, such as engine rotation sound, 15 time-varying sounds, such as engine acceleration sounds, and 15 sudden sounds such as collision of parts. These anomalous sounds were mixed with the normal sounds at signal-to-noise ratios (SNRs) of -10, -5, 0 and 5 dB.

C. Results

We report precision (Prec.), recall (Rec.), and F-measure score (F_1) for the anomaly-detection results from all methods. To evaluate these scores, the threshold value, which maximizes the average score of Prec., Rec., and F_1, was used.

The results are listed in Table I and receiver operating characteristic (ROC) curves are shown in Fig. 3. Overall, the proposed method exhibited the highest F_1 scores, which is a weighted average of Prec. and Rec., compared with the other methods. In addition, Fig. 3 shows that the proposed method improved the TPR compared with the conventional methods

ISBN 978-0-9928626-7-1 © EURASIP 2017

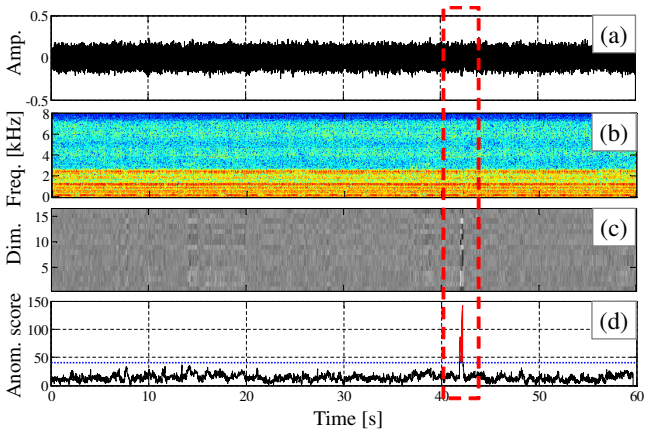

Fig. 4. ASD results on real environment. (a) waveform, (b) spectrogram, (c) acoustic feature, and (d) anomaly score.

even at a low FPR. This may be because the proposed method exhibited a high TPR with a constraint under the low FPR (ρ). These results suggest that the proposed method is more effective than comparison methods.

D. Verification experiment on real environment

We conducted a verification experiment on a real environment. The target equipment was a stereolithography 3D-printer. We collected an actual anomalous sound; a sound caused by collision of the sweeper and formed object. The 3D-printer stopped 5 minutes after this anomalous sound occurred due to the collision.

Normal sounds emitted from the 3D-printer were collected for 30 minutes and used as training data $\mathcal{D}$. We used the same data as various machine operating sounds $\mathcal{O}$ in the objective experiment. The anomalous and normal sounds were recorded at a 16-kHz sampling rate. Since the size of the training data was small, the architecture of $\mathcal{F}$ was as follows: the number of hidden layers was 2 and number of units in each hidden layer was 256. The other conditions were the same as in the objective experiment.

Figure 4 shows the detection results. From the waveform and spectrogram measurements (Figs. 4 (a) and (b), respectively), the anomalous sound could not be clearly identified because the magnitude of the anomalous sound was small. On the other hand, we observed clear changes due to the anomalous sound in the extracted acoustic feature, as shown in Fig. 4 (c). In addition, the anomaly score also increased due to anomalous sound, and the anomalous sound could be identified. This result suggests that the proposed method is effective in identifying anomalous sounds in a real environment.

V. Conclusions

We proposed a method of optimizing an acoustic feature extractor for anomalous-sound detection. By considering outlier-detection-based ASD as a statistical hypothesis test, we defined optimality as an objective function that adopts Neyman-Pearson lemma [18]; the acoustic feature extractor is optimized to extract a set of acoustic features which maximize the true positive rate under an arbitrary false positive rate. The DNN-based feature extractor was optimized with the proposed method by applying the VAE [19]. In the experiments, we found that the F_1 score of the proposed method improved from 0.02 to 0.06 points and could identify anomalous sound in a real environment. Thus, it can be concluded that the proposed method is effective for feature-extractor optimization for ASD.

Acknowledgments: The authors would like to thank NTT DATA Corporation and NTT DATA Engineering Systems Corporation for providing experimental data from a real-environment.

References

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems," *In Proc.* of AVSS, 2007.
[2] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "An Ensemble of Rejecting Classifiers for Anomaly Detection of Audio Events," *In Proc.* of AVSS, 2012.
[3] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust Unsupervised Detection of Human Screams in Noisy Acoustic Environments," *In Proc.* of ICASSP, pp.161–165, 2015.
[4] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio Surveillance of Roads: A System for Detecting Anomalous Sounds," IEEE Trans. ITS, pp.279–288, 2016.
[5] Y. Jiaxing, M. Iwata, T. Kobayashi, M. Murakawa, T. Higuchi, Y. Kubota, Y. Toshiya, and K. Mori, "Statistical Impact-Echo Analysis based on Grassmann Manifold Learning: Its Preliminary Results for Concrete Condition Assessment," *In Proc.* of EWSHM, 2014.
[6] Y. Kubota, Y. E. Jiaxing, M. Iwata, M. Murakawa, and T. Higuchi, "Defect Detection for RC Slab based on Hammering Echo Acoustic Analysis," *In Proc.* of the 30th US-Japan Bridge Engineering Workshop.
[7] V. Chandola, A. Banerjee, and V. Kumar "Anomaly detection: A survey," *ACM Computing Surveys*, 2009.
[8] K. Yamanishi, J. Takeuchi, G. J. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *In Proc.* of ACM SIGKIDD, pp.320–324, 2000.
[9] M. Takimoto, M. Matsugu, and M. Sugiyama, "Visual inspection of precision instruments by least-squares outlier detection," *In Proc.* of DMSS, pp.22–26, 2009.
[10] S. Liu, T. Suzuki and M. Sugiyama, "Support consistency of direct sparse-change learning in Markov networks," *In Proc.* of AAAI, pp.2701–2725, 2015.
[11] T. Ide, A. Khandelwal, and J. Kalagnanam, "Sparse Gaussian Markov Random Field Mixtures for Anomaly Detection," *In Proc.* of ICDM, pp.955–960, 2016.
[12] C. V. Cotton and D. P. W. Ellis, "Spectral vs. Spectrotemporal Features for Acoustic Event Detection," *In Proc.* of WASPAA, 2011.
[13] X. Lu, Y. Tsao, S. Matsuda and C. Hori, "Sparse Representation based on a Bag of Spectral Exemplars for Acoustic Event Detection," *In Proc.* of ICASSP, pp.6255–6259, 2014.
[14] J. Schroder, S. Goetze and J. Anemuller, "Spectro-Temporal Gabor Filterbank Features for Acoustic Event Detection" IEEE Trans. ASLP, 2015.
[15] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *In Proc.* of NIPS, pp.1096–1104, 2009.
[16] M. Espi, M. Fujimoto, K. Kinoshita and T. Nakatani, "Exploiting Spectro-Temporal Locality in Deep Learning based Acoustic Event Detection," EURASIP Journal on Audio, Speech, and Music Processing, 2015.
[17] D. Chakrabarty and M. Elhilali, "Abnormal Sound Event Detection using Temporal Trajectories," *In Proc.* of ICASSP, 2016.
[18] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," Phi. Trans. of the Royal Society, 1933.
[19] D. P. Kingma, and M. Welling, "Auto-Encoding Variational Bayes," *In Proc* of ICLR, 2013.
[20] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition" *Signal Processing Magazine*, 2012.
[21] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *In Proc.* of ICLR, 2015.
[22] A. Krogh and J. A. Hertz, "A Simple Weight Decay Can Improve Generalization," *In Proc.* of NIPS, 1992.

ISBN 978-0-9928626-7-1 © EURASIP 2017