ANOMALY DETECTION BASED ON AN ENSEMBLE OF DEREVERBERATION AND ANOMALOUS SOUND EXTRACTION

Yohei Kawaguchi, Ryo Tanabe, Takashi Endo, Kenji Ichige, and Koichi Hamada

Research and Development Group, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
yohei.kawaguchi.xk@hitachi.com

ABSTRACT

To develop a sound-monitoring system for checking machine health, a method for detecting anomalous sounds is proposed. In real environments such as factories, reverberation and background noise are mixed in an observed signal, so detection performance is degraded. It can be expected that detection performance will be improved by using a front-end algorithm for acoustic signal processing such as dereverberation and denoising. However, any algorithm has pros and cons, so it is not possible to choose the best front-end algorithm only. To solve this problem, the proposed method is based on a front-end ensemble consisting of a blind-dereverberation algorithm and multiple anomalous-sound-extraction algorithms. Experimental results indicate that the proposed method improves detection performance significantly.

Index Terms— machine health monitoring, anomaly detection, ensemble, dereverberation, anomalous sound extraction

1. INTRODUCTION

Sound-monitoring systems for checking a machine's operational condition ("machine health" hereafter) have become more important. Typically, skilled maintenance technicians listen for sounds from machinery and judge the overall condition of a machine from those sounds; however, a shortage of skilled workers has become a serious issue that necessitates an automated system for continuous monitoring of machinery sounds. "Acoustic-scene classification" is a necessary technology for sound-monitoring systems. Especially, as a means of one-class classification for an unsupervised scenario, "anomalous sound detection" is the most applicable for the sound-monitoring systems.

The purpose of this study is to propose a method for detecting anomalous sounds in real environments such as factories. In such environments, reverberation and background noise are mixed in an observed sound signal, so the detection performance is degraded. To solve that problem, the detection method must have robustness against reverberation and background noise. To attain that robustness, two types of architecture are available: a "modular architecture" or an "end-to-end architecture". The modular architecture, which consists of front-end modules and back-end modules working independently, is suitable for anomalous sound detection. In particular, the front-end modules perform acoustic-signal-processing procedures such as "denoising," while the back-end modules perform classification. Although the end-to-end architecture, which is a deep neural network with a high representation power, can be totally optimized by using training data, it is unsuitable for the sound-monitoring systems because a sufficient amount of training data cannot be obtained. In addition, assuring the quality of the end-to-end architecture is a more-difficult process than that for assuring the quality of the modular architecture.

To solve the problem of reverberation and background noise, the proposed architecture applies a front-end ensemble consisting of a blind-dereverberation (BD) algorithm and anomalous-sound-extraction (ASE) algorithms. It can be expected that a dereverberation algorithm and a denoising algorithm can improve the detection performance; however, any algorithm has pros and cons, so it is impossible to choose the "best" front-end algorithm. Thus, as for the proposed method, an ensemble of multiple front-end algorithms is applied. In particular, the BD algorithm proposed by Togami et al. [1], which is suitable for unknown anomalous sounds, is used because it does not rely on training. In addition, multiple ASE algorithms based on non-negative matrix factorization (NMF) [2], non-negative matrix underapproximation (NMU) [3], and non-negative novelty extraction (NNE) [4] are used in parallel, and these algorithms complement each other.

Experimental results indicate that the proposed method improves the detection performance significantly. Also, it is indicated that all of Togami's BD algorithm, the ASE algorithms, and the ensemble architecture improve the detection performance.

2. RELATION TO PRIOR WORK

The three main contributions of this paper are summarized as follows. The first contribution is clarification that an ensemble of multiple front-end procedures is suitable for anomalous sound detection. In the case of past DCASE challenges, some participants applied front-end ensemble approaches and achieved good results for scene-classification tasks in supervised learning scenarios. For example, Han et al. got 2nd place [5] in the DCASE 2017 Task 1 [6], Sakashita et al. got 1st place [7] in the DCASE 2018 Task 1 [8], and Tanabe et al. got 1st place tie [9] in the DCASE 2018 Task 5 [10]. In contrast, as for anomalous sound detection in an unsupervised learning scenario, it is not clear that an ensemble of multiple front-end procedures provides good performance, although anomalous sound detection has been extensively studied [11][12][13][14][15][16][17].

The second contribution is demonstration that dereverberation improves the performance of anomalous sound detection. Although many researches have shown that dereverberation improves accuracy of automatic speech recognition [18][19][20][21], anomalous sound detection based on dereverberation has yet to be studied.

The third contribution is demonstration that ASE algorithms improve the performance of anomalous sound detection. In [4], an ASE algorithm (named NNE) was proposed, and the SNR was improved. However, it was not confirmed that the ASE technique improves the anomaly-detection performance.

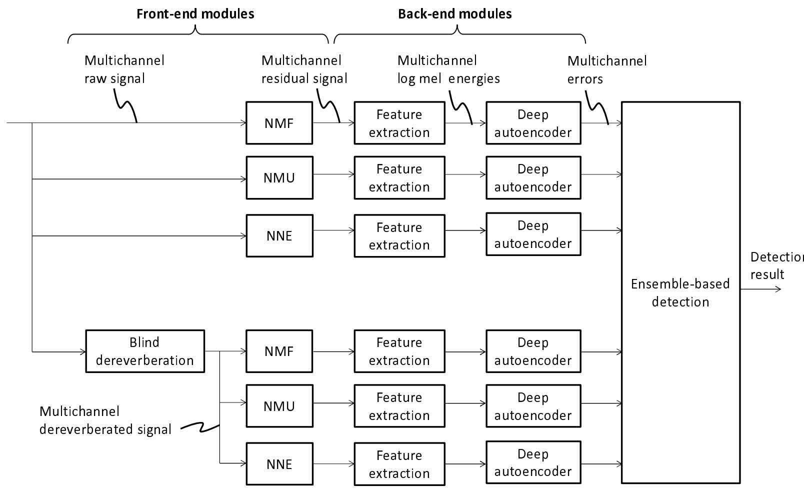

Fig. 1. Proposed architecture

3. PROPOSED METHOD

The architecture of the proposed method is shown in Fig. 1. The system consists of front-end modules, back-end modules, and an ensemble-based detector. The front-end modules consist of a module for BD and modules for ASE.

3.1. Blind Dereverberation (BD)

BD algorithms [1][22] are suitable for the sound-monitoring systems, in which acoustic transfer functions (ATFs) often change, because they do not rely on training. In this study, Togami's BD algorithm [1] is applied because of its usability. It combines multichannel inverse filtering, beamforming, and non-linear reverberation suppression (NRS). It is robust against ATF fluctuations and creates less distortion than NRS alone. The three components are optimally combined from a probabilistic perspective by using a unified likelihood function incorporating a multichannel-probabilistic-source model and a probabilistic-reverberant-transfer-function model. Togami's BD separates the dereverberated signal for each source and the residual reverberation from the M-channel input signal. The dereverberated signal for each source also contains M channels because Togami's BD is a multi-input-multi-output (MIMO) algorithm. By mixing the dereverberated signals of all sources, an M-channel output signal is made and sent to the anomalous sound extractors. Although a dereverberated signal for each source is outputted by Togami's BD, a mixed dereverberated signal is inputted to the next procedure because source separation is not focused on in this study. In the case of DCASE 2018 Challenge Task 5, the authors used not only the dereverberated signal but also the reverberation signal [9], whereas in the case of the proposed method, the reverberation signal is not used because the effect of the dereverberated signal is focused on in this study.

3.2. Anomalous Sound Extraction (ASE)

To solve the background-noise problem, the proposed method uses multiple ASE algorithms. Most algorithms for anomaly detection learn the acoustic model corresponding to the normal sound by using training data in advance. The anomaly-detection algorithms calculate a criterion for determining how different the input signal is from the normal sound [23], and it detect the anomalous sound based on the criterion. In real environments, the input signal includes the background noise belonging to the normal-sound category, so the criterion fluctuates as the noise fluctuates to some extent, and detection performance decreases. Hence, ASE can be considered suitable for anomaly detection in noisy environments.

ASE algorithms are briefly explained as follows. The procedure of ASE consists of two phases, i.e., a training phase and an extraction phase. In the training phase, a normal-sound model consisting of acoustic atoms is estimated from the amplitude spectrogram of the input signal in training data $\boldsymbol{X}$, which is an $N \times T$ matrix, where N is the number of frequency bins, and T is the time index. The training data contains only normal sound; namely, it does not contain anomalous sound. The normal sound is generated from healthy machines or background noise. It is assumed that $\boldsymbol{X}$ can be decomposed into an $N \times K$ matrix called acoustic atoms ($\boldsymbol{W}$) and a $K \times T$ matrix called activations ($\boldsymbol{H}$), where K is the number of acoustic atoms. $\boldsymbol{W}$ can be learned from $\boldsymbol{X}$ in an unsupervised-NMF manner [2]. In the extraction phase, the spectrogram of anomalous sound $\boldsymbol{R}$ is estimated from the input spectrogram in test data $\boldsymbol{X}$ and learned atoms $\boldsymbol{W}$.

As the ASE algorithms used in this study, NMF, NMU, and NNE were chosen. Improvements in SDR yielded by different ASE algorithms are shown in Figure 2. NMF, NMU, and NNE have different correlations between the input SNR and the improved SDR. Therefore, it is expected that these algorithms complement each other.

The NMF-based ASE is the simplest algorithm. It is assumed that $\boldsymbol{X}$ is a linear combination of normal sound $\boldsymbol{WH}$ and anomalous sound $\boldsymbol{R}$ given as follows:

$$\boldsymbol{X} = \boldsymbol{WH} + \boldsymbol{R} \quad \text{s.t.} \quad \boldsymbol{W}, \boldsymbol{H} \geq \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{H}$ is unknown in the extraction phase. $\boldsymbol{H}$ is easily estimated by supervised NMF based on multiplicative updates [2]. The NMF-based ASE estimates $\boldsymbol{R}$ from the estimated $\boldsymbol{H}$, $\hat{\boldsymbol{H}}$, as follows:

$$\hat{\boldsymbol{R}} = \boldsymbol{X} - \boldsymbol{W}\hat{\boldsymbol{H}}, \tag{2}$$

where $\hat{\boldsymbol{R}}$ is estimated $\boldsymbol{R}$. However, the estimated anomalous sound $\hat{\boldsymbol{R}}$ tends to be distorted and vanish under noisy conditions because $\hat{\boldsymbol{R}}$ may have negative values.

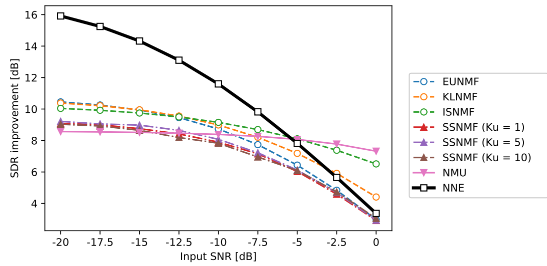

Fig. 2. Improvements in SDR yielded by different ASE algorithms. Number of basis components is 20. The x- and y-axis show input SNR [dB] and average SDR improvement [dB], respectively. (Kawaguchi et al., IWAENC 2018 [4])

NMU [3] is similar to NMF, but $\boldsymbol{R}$ is always kept non-negative because an underapproximation constraint is added. In each iteration, the supervised version of NMU determines $\boldsymbol{h}^T$, which is one row vector of $\boldsymbol{H}$, by solving the following optimization problem:

$$\min_{\boldsymbol{h},\boldsymbol{R}} \frac{1}{2}\|\boldsymbol{R}\|_F^2 \quad \text{s.t.} \quad \boldsymbol{X} = \boldsymbol{w}\boldsymbol{h}^T + \boldsymbol{R}, \quad \boldsymbol{w}, \boldsymbol{h}, \boldsymbol{R} \geq \boldsymbol{0}, \tag{3}$$

where $\boldsymbol{w}$ is the column vector of $\boldsymbol{W}$ corresponding to $\boldsymbol{h}^T$. The NMU-based ASE outputs the estimate of $\boldsymbol{R}$ after a certain number of iterations. The anomalous sound is explicitly modeled as a non-negative matrix, so it can be expected that NMU is superior to NMF. However, in the case of NMU, the error accumulates in each iteration because it is a greedy algorithm, so its extraction performance is not so high.

NNE [4] is a simultaneous-optimization version of NMU revised from NMU, so the demerit of NMU caused by the greedy iterations is eliminated. NMU and NNE are similar in the regard that $\boldsymbol{R}$ is always kept non-negative. NNE-based ASE also outputs estimated $\boldsymbol{R}$ after iterations, where $\boldsymbol{R}$ is estimated by solving the following optimization problem:

$$\min_{\boldsymbol{H},\boldsymbol{R}} \frac{1}{2}\|\boldsymbol{R}\|_F^2 \quad \text{s.t.} \quad \boldsymbol{X} = \boldsymbol{W}\boldsymbol{H} + \boldsymbol{R}, \quad \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R} \geq \boldsymbol{0}. \tag{4}$$

By applying the alternating direction method of multipliers (ADMM), (4) is converted to the following update rules:

$$\boldsymbol{H}_k = \mathcal{S}\left(\boldsymbol{W}_k^T\boldsymbol{M} / \left(\boldsymbol{W}_k^T\boldsymbol{W}_k\right)\right), \tag{5}$$

$$\boldsymbol{R}_k = \mathcal{S}\left(\frac{1}{1+\gamma}\left(\gamma\left(\boldsymbol{X} - \boldsymbol{W}_k\boldsymbol{H}_k\right) + \boldsymbol{\Gamma}_{k-1}\right)\right), \tag{6}$$

$$\boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}_{k-1} + \xi\gamma\left(\boldsymbol{X} - \boldsymbol{W}_k\boldsymbol{H}_k - \boldsymbol{R}_k\right), \tag{7}$$

where k is the index of iterations, γ and ξ are positive constant values, $\boldsymbol{M} = \boldsymbol{X} - \boldsymbol{R}_{k-1} + \gamma\boldsymbol{\Gamma}_{k-1}$, and $\mathcal{S}(\cdot)_{ij} = \max\{(\cdot)_{ij}, 0\}$. The procedure of NNE is shown in Algorithm 1.

3.3. Back-End Modules and Ensemble-Based Detection

In the feature-extraction modules, the log mel energies are calculated from the input signal by 40 mel filters, and a 200-dimensional feature vector is made every five frames. The feature vector is input to a deep autoencoder consisting of three hidden fully connected layers (200-40-40-40-200 units). Rectified linear units (ReLUs) [24] are applied for all hidden layers. The autoencoder corresponding to each front-end module executes a training procedure individually. For training, Adam [25] and batch normalization [26] were applied. For every five frames (τ), namely, every autoencoder (i), and every microphone channel (m), the reconstruction error of the autoencoder, $e_{\tau,i,m}$, is calculated. We expect that even if the autoencoders are replaced with Gaussian-mixture models [11][12], variational autoencoders [15][17], or other anomaly-detection algorithms, the proposed method (namely, an ensemble consisting of a BD algorithm and ASE algorithms) will improve the detection performance in a similar manner to the autoencoders. However, this expectation is beyond the scope of this paper.

Algorithm 1 NNE (supervised version)

Input: $\boldsymbol{X}$ and $\boldsymbol{W}$
Initialize $\boldsymbol{W}_{k-1} = \boldsymbol{W}_{k-1}$
Initialize $\boldsymbol{H}_{k-1}$ by supervised NMF for $\boldsymbol{X}$ and $\boldsymbol{W}$
Initialize $\boldsymbol{R}_{k-1} = \mathcal{S}\left(\boldsymbol{X} - \boldsymbol{W}_{k-1}\boldsymbol{H}_{k-1}\right)$
Initialize $\boldsymbol{\Gamma}_{k-1}$ by random values
repeat
    Update $\boldsymbol{H}_k$ using (5)
    Update $\boldsymbol{R}_k$ using (6)
    Update $\boldsymbol{\Gamma}_k$ using (7)
    $k = k + 1$
until Convergence
Output: $\boldsymbol{H}_k$ and $\boldsymbol{R}_k$

The proposed method evaluates the average ensemble of the reconstruction errors averaged over the whole segment (10 s), all the autoencoders, and all the channels. Although averaging is the simplest way, it is very robust to overfitting. Averaged reconstruction error E is given as

$$E = \sum_{\tau}\sum_{i}\sum_{m} \frac{e_{\tau,i,m} - \mu_i}{\sigma_i}, \tag{8}$$

where μ_i and σ_i are average and standard deviation of $e_{\tau,i,m}$ over τ and m calculated from the training data, respectively. The anomalous sound is detected by thresholding E for each 10-s segment.

4. EXPERIMENTAL RESULTS

We experimentally evaluated the improvement in anomaly detection by applying the proposed method. A circular microphone array consisting of eight microphones (Fig. 4) was arranged around real automated machines consisting of a lot of parts (such as mechanical arms). When the machines repeatedly performed a series of work, the sound of the machines was recorded as 16-bit audio signals sampled at 16 kHz (Fig. 5). The reverberation time was about 400 ms. Recorded data with length of 10 s per segment × 15 segments × 18 machines (= 2700 s) was used as training data, and other data with length of 10 s per segment × 15 segments × 18 machines (= 2700 s) was used as "normal" test data. The same amount of "abnormal" test data was recorded by playing the anomalous sound from a loudspeaker located 1 m from the microphone array in the same environment. The anomalous sound was periodically swept from 0.3 to 7 kHz in a cycle time of 3 s. In the "abnormal" test data, the input SNR was set to -5, -10, and -15 dB by changing the amplitude of the sweep sound. A Hanning window with frame size of 512 and frame shift of 256 was applied, a Euclidean-norm NMF [2] was used for the NMF-based ASE, and number of basis components (K) for ASE was set to 20.

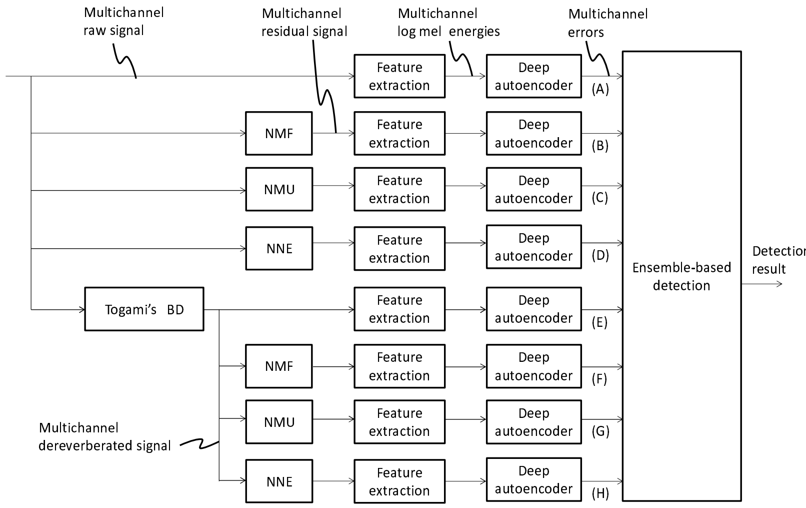

Fig. 3. All ensembles in Section 4

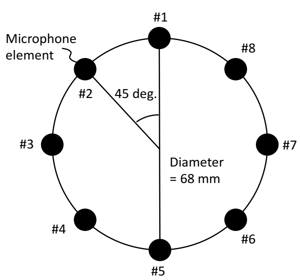

Fig. 4. Circular microphone array

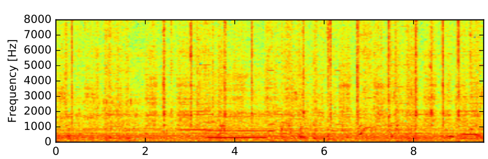

Fig. 5. Spectrogram of a part of the recorded sound. X-axis shows the time in seconds.

"Area under the curve" (AUC) for each autoencoder and each ensemble is listed in Table 1. In the table, each label from (A) to (H) corresponds to autoencoders in Fig. 3. Comparing "w/o BD" with "w/ BD" in Table 1 reveals that BD improves the performance of anomalous sound detection. Moreover, comparing "w/o ASE" with the other ASEs in Table 1 reveals that the ASE algorithms also improve the performance of anomalous sound detection. In addition, when input SNR is -10 dB or -15 dB, the AUCs of the proposed ensemble are higher than those of the best combination of BD and ASE, whereas those of the ensembles (A)-(H), (E)-(H), and (B)-(D) are relatively low. Also, when input SNR is -5 dB, the AUC of the proposed ensemble is about 1, namely, at the same level as those of the best combination of BD and ASE. These results indicate that the front-end algorithms in the proposed ensemble complement each other in noisy cases. It can be considered that the reason that the AUCs of ensemble (A)-(H) are lower than those of the proposed ensemble is that those of (A) and those of (E) are too low.

Table 1. Results of anomalous sound detection (AUC)

BD	ASE	label in Fig. 3	Input SNR -5 dB	-10 dB	-15 dB
w/o BD	w/o ASE	(A)	0.848	0.722	0.654
	NMF	(B)	0.990	0.931	0.697
	NMU	(C)	0.939	0.795	0.675
	NNE	(D)	0.991	0.924	0.760
w/ BD	w/o ASE	(E)	0.886	0.752	0.664
	NMF	(F)	**1.000**	**0.954**	0.785
	NMU	(G)	0.965	0.857	0.732
	NNE	(H)	**1.000**	0.945	**0.813**
ensemble of					
all of (A)-(H)			0.970	0.853	0.703
only BD ((E)-(H))			0.977	0.889	0.744
only ASE ((B)-(D))			0.983	0.880	0.702
proposed ((B)-(D) & (F)-(H))			**0.999**	**0.974**	**0.854**

5. CONCLUSION

A method for detecting anomalous sound was proposed. To solve the problem of reverberation and background noise, the proposed method is based on a front-end ensemble consisting of a BD algorithm and ASE algorithms. The multiple front-end algorithms in the proposed ensemble complement each other. Experimental results indicate that the proposed method improves detection performance significantly. In particular, the BD algorithm, ASE algorithms, and ensemble architecture all individually contribute to improving detection performance.

6. REFERENCES

[1] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1369–1380, July 2013.

[2] D.D. Lee and H.S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Nov. 1999.

[3] M. Tepper and G. Sapiro, "Nonnegative matrix underapproximation for robust multiple model fitting," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nov. 2017, pp. 655–663.

[4] Y. Kawaguchi, T. Endo, K. Ichige, and K. Hamada, "Non-negative novelty extraction: A new non-negativity constraint for NMF," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2018, pp. 256–260.

[5] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov. 2017.

[6] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE2017 challenge setup: Tasks, datasets and baseline system," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov. 2017, pp. 85–92.

[7] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," in *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, Sept. 2018.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov. 2018.

[9] R. Tanabe, T. Endo, Y. Nikaido, K. Ichige, P. Nguyen, Y. Kawaguchi, and K. Hamada, "Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling," in *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, Sept. 2018.

[10] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 challenge - task 5: Monitoring of domestic activities based on multi-channel acoustics," in *arXiv:1807.11246*, July 2018.

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

[12] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 165–168.

[13] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 1996–2000.

[14] 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. International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7.

[15] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," in *Proc. European Signal Processing Conference (EUSIPCO)*, Aug. 2017, pp. 698–702.

[16] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept. 2017.

[17] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 2366–2370.

[18] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB Challenge," in *Proc. REVERB Challenge Workshop*, May 2014.

[19] Y. Tachioka, T. Narita, F.J. Weninger, and S. Watanabe, "Dual system combination approach for various reverberant environments with dereverberation techniques," in *Proc. REVERB Challenge Workshop*, May 2014.

[20] F.J. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J.T. Geiger, B.W. Schuller, and G. Rigoll, "The MERL/MELCO/TUM system for the REVERB Challenge using deep recurrent neural network feature enhancement," in *Proc. REVERB Challenge Workshop*, May 2014.

[21] N. Kanda, R. Ikeshita, S. Horiguchi, Y. Fujita, K. Nagamatsu, X. Wang, V. Manohar, N.E.Y. Soplin, M. Maciejewski, S.-J. Chen, A.S. Subramanian, R. Li, Z. Wang, J. Naradowsky, and G. Sell L.P. Garcia-Perera, "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," in *Proc. 5th International Workshop on Speech Processing in Everyday Environment (CHiME)*, Sept. 2018.

[22] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[23] K. Imoto, "Introduction to acoustic event and scene analysis," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 182–188, May 2018.

[24] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *Proc. 12th IEEE International Conference on Computer Vision (ICCV)*, Sept. 2009, pp. 2146–2153.

[25] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv:1412.6980*, 2014.

[26] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd International Conference on Machine Learning (ICML)*, July 2015, pp. 448–456.