DETECTION OF ANOMALY ACOUSTIC SCENES BASED ON A TEMPORAL DISSIMILARITY MODEL

Tatsuya Komatsu and Reishi Kondo

Data Science Research Laboratories, NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki 211-8666, Japan

ABSTRACT

This paper proposes detection of anomaly acoustic scenes based on a temporal dissimilarity model. The periodicity in the temporal variation of acoustic scenes is first pointed out and then used to build a new stochastic model. In the new model, the temporal variation is expressed by dissimilarity between current and previous acoustic scenes. Anomaly acoustic scenes are detected based on the 24-hour periodic dissimilarity model. Evaluation results using 40-day (1000-hour) data show that the proposed method can detect unknown anomaly acoustic scenes with 82.3% F-measure in 0 dB signal-to-noise-ratio conditions.

Index Terms— anomaly acoustic scene detection, anomaly detection, periodicity, Kullback-Leibler divergence

1. INTRODUCTION

There are serious and growing dangers of terrorism throughout the world. To make our living environment safer, acoustic event detection (AED) for a transient sound and acoustic scene classification (ASC) for a mixture of transient and continuous sounds have been shown to be effective in acoustic monitoring systems for detecting hazardous sounds related to critical incidents such as screaming[1, 2], gunshots[3], and explosion[4]. Most of the conventional AED/ASC methods [5, 6, 7, 8, 9] are based on supervised classification which requires a prior definition of all possible classes and collection of training data consisting of hardly encountered hazardous sounds. To tackle this problem, unsupervised detection methods for anomaly acoustic "events" have been proposed [10, 11, 12]. While the anomaly acoustic "event" has direct relation to a critical incident, it is not easy to detect due to its transient nature. On the other hand, anomaly acoustic "scene" is continuous and easier to detect due to its continuity. However, there is no literature on anomaly acoustic "scenes." Although the anomaly acoustic "scene" may not have direct relations to critical incidents, it is worth detection and further investigation.

Outliers [13, 14, 15] and change points [16, 17, 18, 19] are useful clues for detection of anomaly acoustic scenes. Outlier detection can be treated as one-class classification where anomaly is detected as an outlier of a single normal class. Change point detection models temporal variation of time-series data using dissimilarity between current and previous models, and anomaly is detected as a significant change representing high dissimilarity. However, in the case of anomaly acoustic "scene", definitions of normal and anomaly depend on time. For example, in a station, sudden appearance/disappearance of babble noise with the first/last train is a significant change but normal; silence in the midnight is normal, but is anomaly in the daytime; and laughter in the daytime is normal, but is anomaly in the midnight. In another word, the definitions of normal and anomaly are time-dependent under the same environment. To detect anomaly acoustic "scenes" with time-dependent definitions, conventional outlier/change detection methods need to train and store multiple normal models, the number of which is equal to that of time-dependent definitions. It means that an unrealistic size of memory and a computational cost are needed

However, temporal variation of acoustic scene has a periodicity with a day, a week, a month, and a year to name a few. From a viewpoint of human activities, 24-hour periodicity is most significant. The periodicity makes it possible to establish a statistical model of temporal variation of acoustic scene instead of a huge number of time-dependent multiple normal models. The temporal variation can be interpreted as change of acoustic scene which is modeled with dissimilarity between current and previous acoustic scene models.

This paper proposes detection of anomaly acoustic scenes based on a temporal dissimilarity model. The proposed method models normal temporal variation of time-dependent acoustic scenes using dissimilarity between current and previous acoustic scene models. An anomaly acoustic scene is detected as an "anomaly change" based on 24-hour periodic dissimilarity model.

2. PROPOSED METHOD

The proposed method models temporal variation of time-dependent acoustic scenes using dissimilarity between current and previous acoustic scene models. Anomaly acoustic scenes are detected as an anomaly change of acoustic scene based on 24-hour periodicity of the dissimilarity.

Figs.1 and 2 show a block diagram of the proposed method and relationships among temporal indices $l, m,$

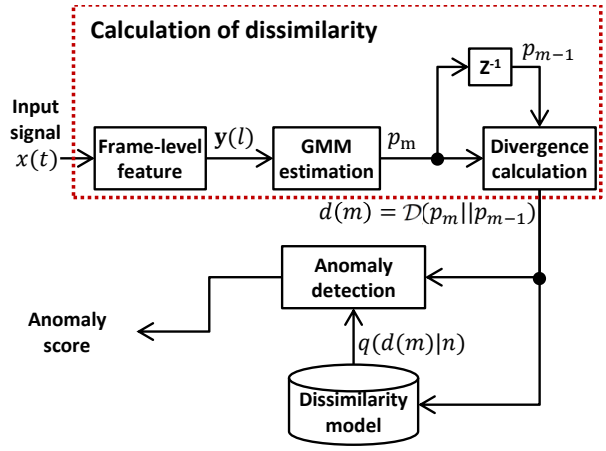


Fig. 1. Block diagram of the proposed method

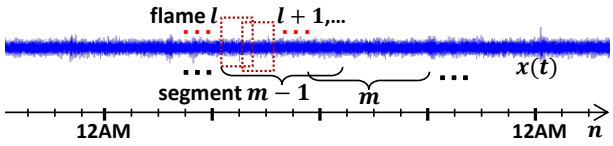


Fig. 2. Relationships among temporal indices l, m, n

and n appeared in this paper. l and m, denote a frame index and a temporal segment index which consists of plural frames respectively, and n represents time of the day with 24-hour periodicity, whose unit is equal to the segment shift of m. To calculate the dissimilarity, an input signal $x(t)$ is transformed to frame-level feature $\mathbf{y}(l)$. Next, a probability density function (PDF) $p(\mathbf{y}(l)|m)$ of $\mathbf{y}(l)$ in the temporal segment m is modeled by a Gaussian mixture model (GMM). Divergence $D(p_m||p_{m-1})$ between $p_m = p(\mathbf{y}|m)$ and $p_{m-1} = p(\mathbf{y}|m-1)$ is calculated and used as temporal dissimilarity $d(m)$. To detect anomaly scene, a temporal dissimilarity $d(m)$ at the corresponding time of day n is modeled as a PDF $q(d(m)|n)$. An anomaly score of $d(m)$ is obtained based on $q(d(m)|n)$.

Generally, an acoustic scene in the real environment has temporal variation. Considering the scene in a station; an acoustic scene begins with silence before the first train and babble noise suddenly appears with the first train. The babble noise continues till the last train at the night changing its statistical property. After the last train, the acoustic scene goes back to silence. This temporal variation of acoustic scene has 24-hour periodicity related to human activities. The proposed method models this temporal variation based on the periodicity.

2.1. Calculation of dissimilarity

This section explains calculation of dissimilarity. First, the proposed method extracts an r dimensional frame-level feature $\mathbf{y}(l)$ from an observed acoustic signal $x(t)$. For the frame-level feature, MFCCs are extracted with a frame length 20 ms and a frame shift 10 ms (50% overlap). Next, a PDF $p_m = p(\mathbf{y}(l)|m)$ of frame-level feature $\mathbf{y}(l)$ in a temporal segment m is calculated; a segment length and a segment shift are set to 5 minutes and 1 minute (80% overlap) respectively. To model $p(\mathbf{y}(l)|m)$, a Gaussian mixture model (GMM) is used:

$$p_m = p(\mathbf{y}(l)|m) = \sum_{k=1}^{K} \pi_{k,m} \mathcal{N}(\mathbf{y}(l)|\boldsymbol{\mu}_{k,m}, \boldsymbol{\Sigma}_{k,m}), \quad (1)$$

where k, K, $\boldsymbol{\mu}_{k,m}$, and $\boldsymbol{\Sigma}_{k,m}$ denote the number of Gaussian components, an index of the components, k-th mean vector of p_m, and k-th covariance matrix of p_m respectively. $p(\mathbf{y}(l)|m)$ represents a statistical property of sounds occurred in the segment m and characterizes an acoustic scene of the segment m. The dissimilarity $d(m)$ in the segment m is calculated using divergence between the PDF $p(\mathbf{y}(l)|m)$ in segment m and the PDF $p(\mathbf{y}(l)|m-1)$ in segment $m-1$,

$$d(m) = \mathcal{D}(p_m||p_{m-1}) \quad (2)$$

There are several alternatives for divergence between GMMs such as Bhattacharyya divergence[20] and Kullback-Leibler(KL) divergence[21]. In this paper, for simplification and reduction of a computational cost, summation of KL divergence between corresponding Gaussian components is used as an approximation of KL-divergence between GMMs,

$$\mathcal{D}_{\mathcal{KL}}(p_m||p_{m-1}) = \sum_{k=1}^{K} \frac{1}{2} \left[\log \frac{|\boldsymbol{\Sigma}_{k,m-1}|}{|\boldsymbol{\Sigma}_{k,m}|} + \mathrm{tr}\left\{ \boldsymbol{\Sigma}_{k,m-1}^{-1} \boldsymbol{\Sigma}_{k,m} \right\} + (\boldsymbol{\mu}_{k,m} - \boldsymbol{\mu}_{k,m-1})^{\mathrm{T}} \boldsymbol{\Sigma}_{k,m-1}^{-1} (\boldsymbol{\mu}_{k,m} - \boldsymbol{\mu}_{k,m-1}) - r \right] \quad (3)$$

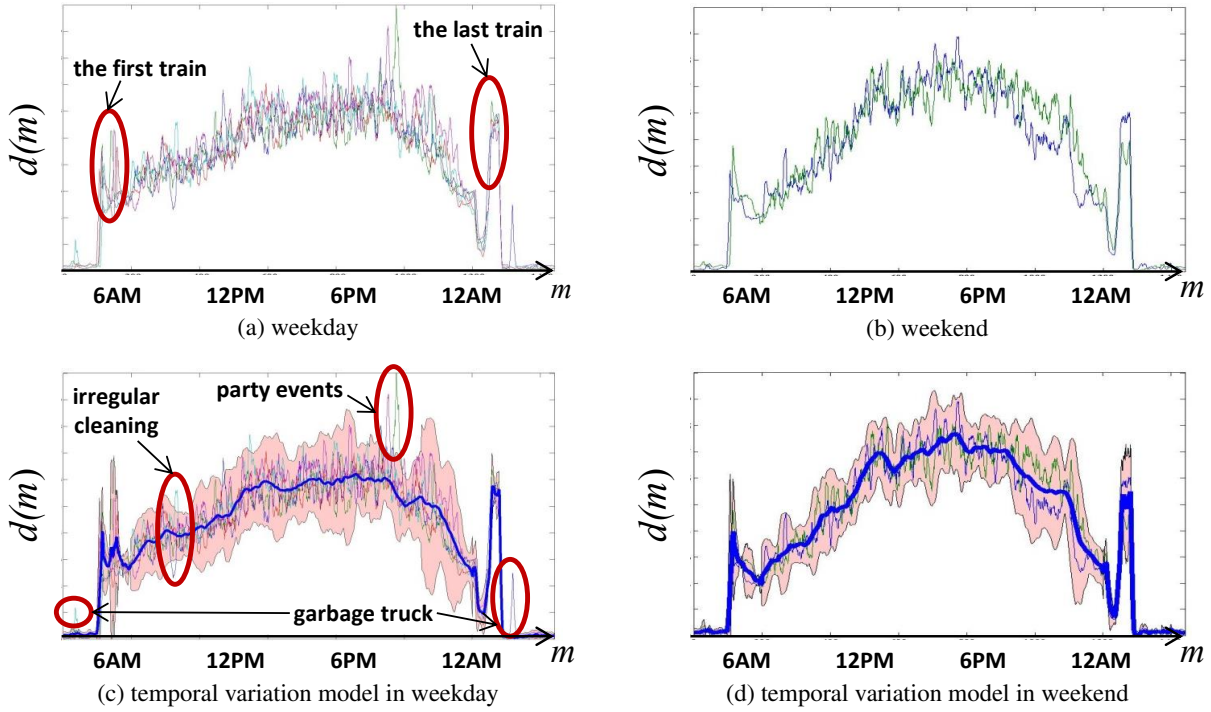
$d(m)$ represents the temporal dissimilarity between acoustic scenes in segment m and $m-1$. Proposed method models temporal variation of acoustic scene with scalar-valued dissimilarity $d(m)$ instead of storing the high dimensional scene models $p(\mathbf{y}|m)$.

2.2. Anomaly detection with temporal dissimilarity model

The proposed method models temporal variation of acoustic scenes using dissimilarity $d(m)$. Generally, acoustic scenes in a real environment have temporal variation with 24-hour periodicity related to human activities. The proposed method models this temporal variation based on the periodicity.

The proposed method models a statistical property of $d(m)$ at each time of day n with a simple Gaussian distribution. n represents time of the day with 24-hour periodicity, whose unit is equal to the segment shift of m. Because there are 1440 minutes in a day and the segment shift m is set to 1 minute, n is an integer of $0-1339$ and m and $m+1440$ represent the same time of the day n. Thus, m and n satisfy

$$m \equiv n \pmod{1440}. \quad (4)$$



(a) weekday (b) weekend

(c) temporal variation model in weekday (d) temporal variation model in weekend

Fig. 3. 24-hour periodicity of dissimilarity

The temporal dissimilarity model $q(d(m)|n)$ at time n is defined as follows:

$$q(d(m)|n) = \mathcal{N}(d(m)|\mu_n, \sigma_n) \tag{5}$$

To estimate a mean μ_n and a variance σ_n, $d(m)$ of corresponding time of the day n,

$$d(n), d(n+1440), d(n+1440\times 2), ... \tag{6}$$

are used. Once $q(d(m)|n)$ is obtained, an anomaly score $e(m_*)$of an acoustic scene in a new temporal segment m_* is calculated based on the probability

$$e(m_*) = q\left(d(m_*)|n_*\right) = \mathcal{N}(d(m_*)|\mu_{n_*}, \sigma_{n_*}) \tag{7}$$

where n_* represents the corresponding time of the day of m_* and satisfies

$$m_* \equiv n_* \pmod{1440}. \tag{8}$$

Using the anomaly score $e(m_*)$, the segment m_* is determined as anomaly acoustic scene based on the following criteria:

$$d(m_*) = \begin{cases} \text{anomaly} & (e(m_*) < \alpha)\,, \\ \text{normal} & (e(m_*) \geq \alpha)\,. \end{cases} \tag{9}$$

Table 1. Parameter setting for the evaluation.

Parameter	Value
data length	40 days (1000 hours)
sampling rate	48 kHz
frame-level feature	MFCC, Δ, $\Delta\Delta$
order of MFCC	13
frame length	20 ms
frame shift	10 ms (50% overlap)
number of GMM components	128
segment for GMM	5 minutes
segment shift	1 minute (80% overlap)

3. EXPERIMENTS

For evaluation, 40-day data is recorded consecutively at a subway station. Table 1 shows a parameter setting used in the evaluation.

3.1. Qualitative evaluation

Fig. 3 shows temporal variation of dissimilarity $d(m)$ extracted from one week data. Figs. 3-(a) and 3-(b) represent $d(m)$ in weekday and weekend respectively. Figs. 3-(c) and 3-(d) represent show mean and standard deviation of temporal dissimilarity model $q(d|n)$ in weekday and weekend respec-

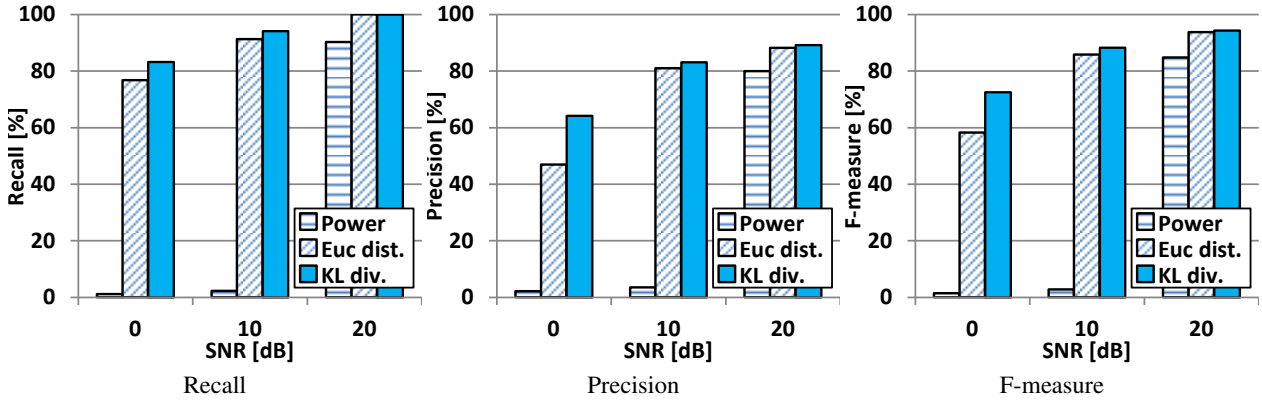


Fig. 4. Experimental results

tively. Each line represents $d(m)$ at corresponding time of the day. Blue bold line and red area show mean and standard deviation respectively. Some characteristics of $d(m)$ are indicated from Fig. 3:

- The temporal variation of $d(m)$ has a similar behavior every day.
- The temporal variation of $d(m)$ has different characteristics in weekdays and the weekend (Fig. 3-(a) and 3-(b)).
- $d(m)$ varies in accordance with human activities: silence of midnight, the first/last train, babble noise in the daytime, and so on.

The temporal variation of acoustic scene in the real environment is modeled properly by the dissimilarity $d(m)$.

In addition, outliers shown in 3-(b) are checked by artificially and it is confirmed that all the outliers show unusual scenes of the environment; appearance of a garbage truck in midnight, irregular operations for construction, cleaning, and some party events.

3.2. Quantitative evaluation

The detection performance of the proposed method for unknown anomaly sounds is evaluated with 40-day recorded data. The temporal dissimilarity model $q(d|n)$ is trained from data of the first week. Evaluation data is made by adding anomaly sounds, which consists of shouts of human crowd, to the original data to simulate panic of human crowd in the subway station caused by some critical incidents such as a riot and terrorism. Sounds are used from Sound Ideas Series 6000 General Sound Effects Library [22]. Duration of each added sound is several tens of seconds. Each sound is added at random temporal positions with three SNRs (signal-to-noise ratio): 0 dB, 10 dB, and 20 dB. Detection performance is evaluated with segment-base recall, precision, F-measure in a day.

With evaluation, two other alternatives of divergence calculation for dissimilarity $d(p_m||p_{m-1})$ are tested. One is the difference of average power of signal between segment m and $m-1$. Another one is a Euclid distance between Gaussian super vectors, which are made by concatenating mean vectors of each Gaussian component of $p(d|m)$ and $p(d|m-1)$:

$$\mathcal{D}_{\mathcal{EUC}}\left(p_m||p_{m-1}\right) = \sum_{k=1}^{K} |\mu_{k,m} - \mu_{k,m-1}|^2. \qquad (10)$$

Fig. 4 shows evaluation results. At 20 dB, all dissimilarity can detect anomaly acoustic scenes with high performances. At low SNRs, temporal variation of power does not work at all. Dissimilarity with the KL divergence shows low degradation of performance at low SNRs compare to the Euclid distance. This is because KL divergence can express differences of not only means but also covariance, so the KL divergence out performs the Euclid distance at low SNRs. It should be noted that the evaluation data includes added anomaly scenes and real anomaly scenes in the original data. The real anomaly scenes are treated as normal scene in the evaluation metrics and cause inevitable degradation of precision.

4. CONCLUSIONS

This paper has proposed detection of anomaly acoustic scenes based on a temporal dissimilarity model. The periodicity in the temporal variation of acoustic scenes has been first pointed out and then used to build a new stochastic model. The temporal variation is expressed by dissimilarity between current and previous acoustic scenes. Anomaly acoustic scenes are detected based on the 24-hour periodic dissimilarity model. Evaluation using 40-day (1000-hour) data have shown that the proposed method can detect unknown anomaly acoustic scenes with 82.3% F-measure in 0 dB signal-to-noise-ratio conditions.

5. REFERENCES

[1] Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, Fabio Antonacci, and Augusto Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. IEEE, 2007, pp. 21–26.

[2] Tatsuya Komatsu, Yuzo Senda, and Reishi Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2259–2263.

[3] Chloé Clavel, Thibaut Ehrette, and Gaël Richard, "Events detection for an audio-based surveillance system," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 1306–1309.

[4] Yang Zhang, Nasser M Nasrabadi, and Mark Hasegawa-Johnson, "Multichannel transient acoustic signal classification using task-driven dictionary with joint sparsity and beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1866–1870.

[5] Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Jyri Huopaniemi, and Timo Sorsa, "Computational auditory scene recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 2, pp. II–1941.

[6] Antti Eronen, Juha Tuomi, Anssi Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi, "Audio-based context awareness-acoustic modeling and perceptual evaluation," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP). 2003 IEEE International Conference on*. IEEE, 2003, vol. 5, pp. V–529.

[7] Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, Mark D. Plumbley, Peter Foster, Emmanouil Benetos, and Mathieu Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, Tampere University of Technology. Department of Signal Processing, 2016.

[8] Hamid Eghbal-Zadeh, Bernhard Lehner, Matthias Dorfer, and Gerhard Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," Tech. Rep., DCASE2016 Challenge, September 2016.

[9] Victor Bisot, Romain Serizel, Slim Essid, and Ga?l Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," Tech. Rep., DCASE2016 Challenge, September 2016.

[10] Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 713–719, 2011.

[11] Erik Marchi, Fabio Vesperini, Florian Eyben, Stefano Squartini, and Björn Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1996–2000.

[12] Debmalya Chakrabarty and Mounya Elhilali, "Abnormal sound event detection using temporal trajectories mixtures," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 216–220.

[13] Lionel Tarassenko, P Hayton, N Cerneaz, and M Brady, "Novelty detection for the identification of masses in mammograms," in *Artificial Neural Networks, 1995., Fourth International Conference on*. IET, 1995, pp. 442–447.

[14] Jonathan Foote, "Automatic audio segmentation using a measure of audio novelty," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*. IEEE, 2000, vol. 1, pp. 452–455.

[15] Markos Markou and Sameer Singh, "Novelty detection: a review?part 1: statistical approaches," *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[16] Michèle Basseville, Igor V Nikiforov, et al., *Detection of abrupt changes: theory and application*, vol. 104, Prentice Hall Englewood Cliffs, 1993.

[17] Claude Barras, Xuan Zhu, Sylvain Meignier, and J-L Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[18] Yumi Ono, Yoshifumi Onishi, Takafumi Koshinaka, Soichiro Takata, and Osamu Hoshuyama, "Anomaly detection of motors with feature emphasis using only normal sounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2800–2804.

[19] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, vol. 43, pp. 72–83, 2013.

[20] Peder A Olsen and John R Hershey, "Bhattacharyya error and divergence using variational importance sampling.," in *INTERSPEECH*, 2007, pp. 46–49.

[21] John R Hershey and Peder A Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Acoustics, Speech and Signal Processing (ICASSP), 2007 IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–317.

[22] Sound Ideas, "Series 6000 general sound effects library," http://www.sound-ideas.com/sound-effects/series-6000-sound-effects-library.html.