Adaptive training of vibration-based anomaly detector for wind turbine condition monitoring

Takanori Hasegawa[1], Jun Ogata[2], Masahiro Murakawa[3], Tetsunori Kobayashi[4] and Tetsuji Ogawa[5]

[1,4,5] *Dept. of Communications and Computer Engineering, Waseda University,*
Tokyo, 169-8555, JAPAN

[1,2,3,5] *National Institue of Advanced Industrial Science and Technology (AIST),*
Ibaraki 305-8560, JAPAN

ABSTRACT

Adaptive training of a vibration-based anomaly detector for wind turbine condition monitoring system (CMS) is carried out to achieve high-performance detection from the early stages of monitoring. Machine learning-based wind turbine CMSs are required to collect large-scale data to yield reliable predictions. Existing studies in this area have postulated that both data for training a monitoring system and those during the operation of the system are obtained from identical devices. In addition, constant monitoring of data is desirable, but in practice, the data can be observed periodically (e.g., several tens of seconds of data are observed every two hours). In this case, collecting sufficient data is time consuming, making it difficult to conduct accurate predictions at the early stage of the CMS operation. To address this problem, a small amount of vibration data observed at a target wind turbine is utilized to adapt the anomaly detector that is trained on relatively large-scale vibration signals obtained from other wind turbines. In the present study, maximum a posteriori (MAP) adaptation is applied to a Gaussian mixture model (GMM)-based anomaly detector. Experimental comparisons using vibration data from the gearbox in the experimental environment and those used in the wind turbine demonstrated that MAP-based GMM adaptation yielded an improvement in anomaly detection accuracy even when only a small amount of data is observed at the target gearbox.

1. INTRODUCTION

An unexpected arrest of massive infrastructures of renewable energy sources such as wind turbines inflicts enormous damages on society. It is important to reduce downtime of a whole plant by detecting failures of individual machinery at their presage stages such that maintenance can be carried out in a

timely manner. Condition monitoring system (CMS) plays an important role in establishing such condition-based maintenance. Therefore, the development of accurate anomaly or fault detection methods that utilize machine learning technologies is required to provide autonomous online CMSs.

Many attempts have been made to employ machine learning technologies in failure diagnosis or anomaly detection of rotating machinery. Most techniques have been developed under the assumption that the vibration signals observed from the devices follow a single Gaussian distribution (Stander, 2002; Bartkowiak & Zimroz, 2011). In contrast, Gaussian mixture models (GMMs) are used to represent distributions of the vibration signals for anomaly detection of wind turbine components (Ogata & Murakawa, 2016). In actual wind turbines, the vibration data of the components have a wide variety of characteristics even when collected in normal (healthy) conditions due to the various operating states of components and environmental conditions. The effect of such variation is successfully modeled by using GMMs (Ogata & Murakawa, 2016).

An alternative failure diagnosis approach using machine learning has been made to predict the remaining useful life (RUL) of machinery. RUL prediction has been formulated using a regression model, which predicts an RUL directly from the feature parameters (Gebraeel & Lawley, 2008; Guclu, Yilboga, Eker, Camci, & Jennions, 2010; Wang, 2012; Deutsch & He, 2016), and a state transition model, in which the state transits from a normal state to an abnormal state (Camci & Chinnam, 2010; Kim, Tan, Mathew, & Choi, 2012; Medjaher, Tobon-Mejia, & Zerhouni, 2012; Liu, Zuo, & Zhang, 2014). Ideal RUL estimation can specify the status of the machinery or the time to breakdown, making it possible to provide maintenance appropriate to the machinery's situation. Rotating machinery, however, has a large variety of anomalies. It is inherently difficult to accurately predict RULs for unexpected, unknown, or abnormal situations, indicating a need

for large-scale data. In addition, a few failures can be observed in a massive power-generating system, making big-data collection unrealistic. We therefore focus on methods that do not require anomaly data.

Machine learning-based anomaly detection requires data collection to train the detector. In this case, both the development and operation of monitoring systems being made for the individual devices indicate a high accuracy of anomaly detection. Existing anomaly detection systems including the one using GMM practically have been assumed to be constructed for each device. However, since such massive equipment is composed of an enormous number of devices, the operational cost of vibration sensors cannot be negligible. Current anomaly detection systems take a practical approach in which the vibration data are not constantly recorded but periodically, e.g., every few hours. It requires a long time to collect enough data to achieve a higher performance in detection systems for the individual devices. In this work, a method to transfer an existing anomaly detector to another device with similar characteristics is introduced to boost the accuracy at an earlier stage of the CMS operation.

We assume that both types of gearboxes and their places to be installed are different among the training and run-time stages in the anomaly detector, and attempt to adaptively refine the existing GMM-based anomaly detector using small amounts of data obtained in run-time. Specifically, maximum a posteriori (MAP) adaptation (Lee & Gauvain, 1993), which has been effective in adaptive training for GMM, is applied. The knowledge obtained from the present work could be useful in the efficient operation (i.e., reduction in operational cost) of robust CMS against real environment situations.

The rest of the present paper is organized as follows. Section 2 briefly reviews the anomaly detection method using GMM. Section 3 describes the adaptive training method used for GMM-based anomaly detectors. Section 4 investigates the effectiveness of adaptive training of the anomaly detector using the vibration signals obtained from the gearboxes. Finally, a summary is presented in Section 5.

2. GMM-BASED ANOMALY DETECTION SYSTEM

Figure 1 illustrates an overview of the anomaly detection system used for condition monitoring of a rotary device (Ogata & Murakawa, 2016). For training a GMM-based anomaly detector, feature vectors are extracted from vibration signals collected under the condition that the device is operating normally. GMMs are then trained on such a vector space to construct a "normal status" model. At a run-time stage, input vibration signals are transformed into feature vectors in the same manner as those in the training phase. A likelihood of the input vector for the normal status model is then calculated to measure the anomaly of the device operating. In this case, lower likelihoods indicate more abnormal situations of

the device. Thresholding the likelihood is carried out to judge whether the target device is normally or abnormally operating.

Section 2.1 describes feature extraction, in which local autocorrelations on time-frequency patterns are extracted, and Sect. 2.2 describes a method of developing a GMM-based anomaly detection system and an algorithm for detecting anomaly from the input signals.

2.1. Feature extraction method for rotary devices

This subsection describes a method of extracting features from vibration signals for anomaly detection. In CMS, feature representations for a device's health status can affect the performance in anomaly detection. Other attributes such as temperature can be combined with the vibration-derived features to improve the accuracy. It, however, should be noted that detail investigation on effective feature representations is not the focus of this work.

2.1.1. Time series analysis using sliding window

Sliding windows have been employed in analyzing varying signals with a trade–off between temporal and frequency resolution. First, assume that time series data with its length of T are given as:

$$\mathcal{D} = \left\{ \boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \cdots, \boldsymbol{\xi}^{(T)} \right\}.$$

Then, w-dimensional vectors are extracted from the original time series $\mathcal{D}$ using the sliding window with the size of w as:

$$\boldsymbol{X}_1 = (\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \cdots, \boldsymbol{\xi}^{(w)}),$$
$$\boldsymbol{X}_2 = (\boldsymbol{\xi}^{(2)}, \boldsymbol{\xi}^{(3)}, \cdots, \boldsymbol{\xi}^{(w+1)}),$$
$$\cdots$$
$$\boldsymbol{X}_i = (\boldsymbol{\xi}^{(i)}, \boldsymbol{\xi}^{(i+1)}, \cdots, \boldsymbol{\xi}^{(w+i-1)}),$$
$$\cdots$$

where w is determined considering the aforementioned trade off. A Hamming window is applied to a sub sequence $\boldsymbol{X}_t$. Windowed sub sequences are then taken as the inputs to the subsequent feature extraction, where temporal units of individual sub sequences are referred to as "frames."

2.1.2. Fourier local autocorrelation (FLAC)

The time-frequency feature representations are calculated from vibration signals using Fourier local autocorrelation (FLAC) (Ye, Kobayashi, & Higuchi, 2010a). FLAC has been shown to be effective in analyzing not only acoustic signals (Ye et al., 2010b, 2012) but also vibration signals (Ogata & Murakawa, 2016). FLAC is aimed at extracting dynamic transition information on the time-frequency domain and is
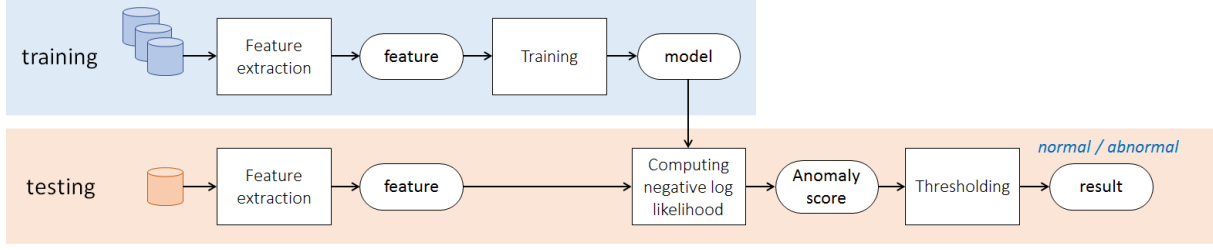
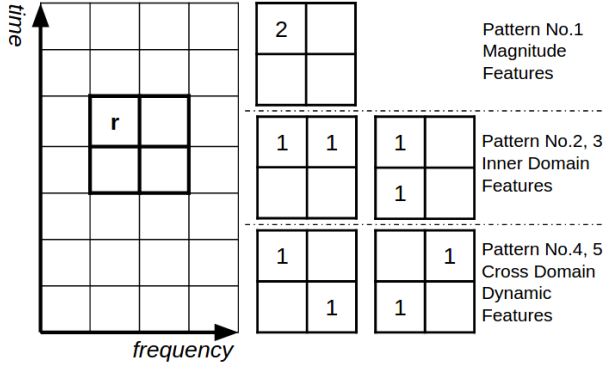Figure 1. Schematic diagram of anomaly detection

Figure 2. Autocorrelation patterns in FLAC feature extraction

obtained by calculating the local autocorrelation for individual time-frequency bins.

In FLAC processing, a series of vibration signals (i.e., the blocked sequence yielded in Sect. 2.1.1) is first transformed to a spectrogram with a short-time Fourier transform. The complex spectrogram $f(\boldsymbol{r})$ at a time-frequency bin $\boldsymbol{r} = (t, v)$ develops the local autocorrelation function utilizing the complex values as:

$$x_{t,v}(\boldsymbol{a}) = f^*(\boldsymbol{r})f(\boldsymbol{r} + \boldsymbol{a}), \qquad (1)$$

where $\boldsymbol{a}$ denotes a displacement vector that represents local neighborhoods and f^* denotes the complex conjugate of f. Figure 2 illustrates the combination pattern of $\boldsymbol{r}$ and $\boldsymbol{r} + \boldsymbol{a}$. In the present study, the displacement vector $\boldsymbol{a}$ is limited to a 2×2 region on the time-frequency plane. In addition, five masking patterns described in Fig. 2 are applied to individual time-frequency bins, expanding each component to a five-dimensional vector. For each frame, the five-dimensional parameters are concatenated across all time-frequency components, yielding high-dimensional vectors. Mel-filterbank analysis is conducted before FLAC extraction to reduce the dimensionality of resulting features. Since local autocorrelation is calculated directly from complex values, the resulting features take dynamics of magnitudes as well as those of phases into account, yielding robustness against phase shift.

2.1.3. Normalization and dimensionality reduction

The difference in scales amongst individual dimensions is decreased to make anomaly detection systems more reliable. In the present study, two scale normalization methods were employed as follows:

$$\tilde{x}_t^d = \frac{x_t^d - \mu^d}{\sigma^d} \qquad (2)$$

$$\tilde{x}_t^d = \frac{x_t^d - v^{d,\min}}{v^{d,\max} - v^{d,\min}} \qquad (3)$$

where x_t^d denotes the d-th component of a feature vector; $\tilde{x}_t^d$ denotes the corresponding scaled component; μ^d and σ^d denote the mean and standard deviation for d-th components, respectively; and $v^{d,min}$ and $v^{d,max}$ denote the minimum and maximum value for d-th components. The former (in Eq. 2) scales the components $\{x_t^d\}$ such that the mean and variance of scaled components $\{\tilde{x}_t^d\}$ would take zero and one, respectively. The latter (in Eq. 3) scales $\{x_t^d\}$ such that the minimum and maximum values of $\{\tilde{x}_t^d\}$ would take zero and one, respectively. Hereafter, the latter method is focused on because the better accuracy was obtained in the preliminary experiment.

In addition, the dimensionality of the feature vector is reduced with principal component analysis (PCA) (Wold, Esbensen, & Geladi, 1987; Jolliffe, 2002) because the dimensionality of FLAC features are high to represent the data distribution using GMMs. The dimensionality-reduced vectors are taken as the inputs for the GMM-based anomaly detector.

2.2. Anomaly detection using GMMs

Most existing techniques for machine learning-based failure diagnosis have exploited single Gaussian distributions to represent vibration signals obtained from rotary devices. The distribution of data observed, however, varies due to the difference in weather conditions even when the device runs under the normal status. An attempt, therefore, is made to employ a GMM for modeling the normal status of the device.

3

The probability density function of GMM is represented as:

$$p(\boldsymbol{x}_t|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{4}$$

where $\boldsymbol{x}_t$ denotes a feature vector for partial time series; K denotes the number of Gaussians; $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ denotes a parameter set of a GMM; π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ denote the mixture weight, mean vector, and full covariance matrix for the k-th component, respectively; and $\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes k-th Gaussian distribution. The parameter set $\boldsymbol{\theta}$ is estimated using an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) on large-scale data of the devices operating normally.

An anomaly score of the input $\boldsymbol{x}_t$ is defined using a logarithmic likelihood as:

$$a(\boldsymbol{x}_t) = -\log p(\boldsymbol{x}_t|\boldsymbol{\theta}) \tag{5}$$

This negative logarithmic likelihood of the normal model takes lower values for the inputs with the normal status and higher values for the abnormal inputs. In this system, the input is regarded as the abnormal status when the anomaly score $a(\boldsymbol{x}_t)$ exceeds a predefined threshold.

3. ADAPTIVE TRAINING OF GMM-BASED ANOMALY DETECTOR

The present study provides a technique to transfer an existing anomaly detector to other similar devices. A mismatch between the vibration data used in developing a type of general-purpose anomaly detector and those observed at run-time can induce the errors in detection. To suppress the effect of such a mismatch, an attempt is made to adapt a GMM-based anomaly detector trained on relatively large-scale data, using a small amount of data obtained from the target device. In the present study, maximum a posteriori (MAP) adaptation is employed to exploit prior information obtained from the other device. This approach makes it possible to boost the accuracy of detecting anomaly even at an early stage of monitoring.

3.1. MAP adaptation

MAP adaptation has been frequently applied to GMM-based prediction systems for reducing mismatches in domains between the development and run-time data. For example, in automatic speech recognition, MAP adaptation has been shown to be effective in compensating the effect of the difference in speakers on the accuracy of GMM/HMM-based acoustic models (Lee, Juang, & Lin, 1991; Gauvain & Lee, 1994).

In MAP adaptation, a new parameter is estimated so as to maximize $p(\boldsymbol{\theta}|\{\boldsymbol{x}_t\})$, which is the posterior probability of the data observed during run-time being generated from the model. In the present study, only the mean vector out of the

GMM parameters is updated due to a small amount of adaptation data. The mean vector of the k-th Gaussian component is estimated into $\hat{\boldsymbol{\mu}}_k$ by weighted averaging of a mean of the pre-trained model $\boldsymbol{\mu}_k^{(0)}$ and a mean calculated on data observed $\{\boldsymbol{x}_t\}$ as:

$$\hat{\boldsymbol{\mu}}_k^{\mathrm{MAP}} = \frac{\tau_k \boldsymbol{\mu}_k^{(0)} + \sum_t p(k|\boldsymbol{x}_t)\boldsymbol{x}_t}{\tau_k + \sum_t p(k|\boldsymbol{x}_t)}, \tag{6}$$

where $\boldsymbol{\mu}_k^{(0)}$ denotes a mean vector of the k-th component of the pre-trained normal status GMM; τ_k denotes a contribution weight for the pre-trained model; $\boldsymbol{x}_t$ denotes an input observed during run-time; and $p(k|\boldsymbol{x}_t)$ denotes a posterior probability of $\boldsymbol{x}_t$ being generated from the k-th component. MAP adaptation explicitly exploits prior information, which are accumulated in the other devices in the present study, through the contribution weight τ_k. The use of statistics on prior information (i.e., $\tau_k \boldsymbol{\mu}_k^{(0)}$ in Eq. 6), can contribute to the improvement in robustness of the parameter estimates against the small size of data.

3.2. ML training

Compared to MAP adaptation, a normal status GMM is trained on data collected from only the target device (i.e., without any prior information obtained from other devices) in maximum likelihood (ML) manner. Here, ML estimates are obtained such that the likelihood $p(\{\boldsymbol{x}_t\}|\boldsymbol{\theta})$ would be maximized. The ML estimate of the mean vector is represented as:

$$\hat{\boldsymbol{\mu}}_k^{\mathrm{ML}} = \frac{\sum_t p(k|\boldsymbol{x}_t)\boldsymbol{x}_t}{\sum_t p(k|\boldsymbol{x}_t)}. \tag{7}$$

In this case, a small amount of data obtained during run-time implies lower reliability of the posterior probability $p(k|\boldsymbol{x}_t)$, indicating that large-scale data are required during training.

4. EXPERIMENTS

Experimental comparisons using vibration signals collected from different gearboxes were carried out to validate the impact of adapting an existing GMM-based anomaly detector to vibration signals obtained from a "target" gearbox. In order to demonstrate an advantage of MAP adaptation using prior information from other similar devices, the two anomaly detection systems were evaluated as follows:

- **ML-train** : a system trained on data obtained from only the target device in maximum likelihood manner.

- **MAP-adapt** : a system in which a prior model is adapted to data obtained from the target device using MAP adaptation.

Section 4.1 describes the data sets used. Section 4.2 describes the criterion for evaluating the accuracy of vibration-based anomaly detectors. In Sections 4.3 and 4.4, the accuracy of

4

an anomaly detector developed on the basis of ML training and that using MAP adaptation were evaluated for various lengths of training/adaptation data to clarify the effectiveness of MAP adaptation at an early stage of collecting data.

4.1. Vibration materials

Two types of vibration data were used; NREL data, which were recorded in the experimental environment, and HSG data, which were obtained from the actual wind turbines. In the present experiment, NREL data were used for development of a prior model and HSG data were used for adaptation and validation. In addition, these data were sampled at different sampling rates. The data with the higher sampling rate were down-sampled to the lower sampling rate.

4.1.1. NREL dataset

The National Renewable Energy Laboratory (NREL) has provided the "wind turbine gearbox condition monitoring vibration analysis benchmarking datasets" (Sheng, 2014) for developing novel technologies on the diagnosis analysis of rotary devices for wind turbine generator systems. Vibration signals were collected in an NREL dynamometer test facility (DTF) under two conditions; a normal operation and an abnormal operation (induced by oil-loss). The turbine tested is a three-bladed, upwind turbine with a rated power of 750 kW. The turbine generator operates at 1800 rpm and 1200 rpm. Eight accelerometers were mounted on the bottom (i.e., six o'clock position) of the ring gear radial to obtain vibration signals in the normal and abnormal status. The vibration data were sampled at 40 kHz. The data was assigned either a "healthy" or "damaged" status label. Each class includes ten files, each of which contains ten seconds of vibration signals.

4.1.2. High speed gear (HSG) dataset

"High speed gear (HSG) dataset" was measured by Eric Bechhoefer and provided through data-acoustics.com (Bechhoefer, 2014). Vibration signals were collected under three conditions; one abnormal condition in which the device was stopped one week after a failure on a pinion gear was found, and two normal conditions in which no known failures were found. The target wind turbine is a three-bladed, upwind turbine with a rated power of 3 MW. The data were sampled at 97.656 kHz. An accelerometer was installed to sense the signals. The data were assigned "case1," "case2," and "case3" labels. The "case1" data were collected under the abnormal condition, and the others were obtained under the normal conditions. The "case1," "case2," and "case3" data included eleven, seven and six files, respectively. Each file contained six seconds of vibration signals.
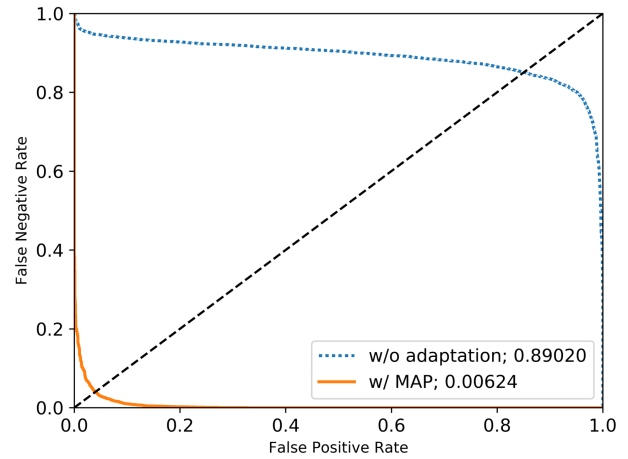

Figure 3. ROC curves yielded from anomaly detection systems with and without MAP adaptation to calculate AUC values. This figure was drawn for the case in which the number of Gaussians was 512 and dimensionality of feature vectors was reduced to two using PCA. Numbers in legend express best AUC values for individual systems.

Table 1. Data set used and their lengths

		data length [frame]	
	data set	normal	abnormal
training	NREL	59892	0
adaptation	HSG	4130 (case2)	0
validation	HSG	3540 (case3)	6490 (case1)

4.2. Evaluation criterion

The anomaly detection system developed judges an input to be an abnormal status if the anomaly score described in Eq. 5 exceeds the predefined threshold. In this case, the accuracy of the system has a trade-off between the false positive rate, which represents a ratio of misjudging abnormal data to be normal status, and false negative rate, which represents a ratio of misjudging normal data to be abnormal status. To consider such trade-off in evaluation, receiver operating characteristic (ROC) curves are exploited (Lusted, 1971; Goodenough, Rossmann, & Lusted, 1974; Metz, 1978). Figure 3 shows an example of ROC curves. The horizontal and vertical axes are the false positive and false negative rate, respectively. Since the purpose of anomaly detection is to reduce both the false positive and false negative rate, the area under the curve (AUC) (Hanley & Mcneil, 1982) should be small to achieve better accuracy in anomaly detection.

4.3. Experimental setup

Table 1 lists the size of data used for training, adaptation, and validation. HSG data, which were originally sampled at 97.656 kHz, were down-sampled to 40 kHz to match the sampling rates of two datasets. NREL data with the "healthy"

status label were used for developing the prior GMM and PCA projection matrix. In the present study, HSG data were considered to be vibration signals from the target gearbox and used during adaptation and validation. Specifically, the "case2" normal status data from HSG data were used to estimate the GMM parameters in maximum likelihood scheme or update the prior GMM parameters using MAP adaptation. The "case1" abnormal status data and "case3" normal status data were used for testing. The "case2" data included 7 out of 13 files. The present experiment investigates the AUCs as a function of adaptation data sizes (one file to seven files) to demonstrate that MAP adaptation performs better than ML training, particularly at the earlier stage of collecting data.

Here, anomaly scores were calculated as described in Eq. 5 and thresholded frame-by-frame to draw ROC curves and calculate AUCs for all the validation data. In addition, tuning parameters in the system developed were determined from the preliminary experiments. The sliding window length used was 0.1s (i.e., w was 4000 samples); the number of mel-filterbanks was 15, yielding a 75-dimensional (15×5) FLAC vector and reducing the dimensionality into two using PCA for both ML training and MAP adaptation; τ_k in Eq. 6 was set to four; and the number of Gaussians in MAP adaptation and ML training were respectively set to 512 and 64. In MAP adaptation, the number of Gaussian components in a prior GMM can be large because the prior GMM to be adapted was trained on large-scale data while in ML training, all the GMM parameters were trained on small amount of data. This is the reason why the optimal number of Gaussians varies between MAP adaptation and ML training.

4.4. Experimental results

Figure 4 and Table 2 show the AUC values calculated from two anomaly detection systems (i.e., **ML-train** and **MAP-adapt**) as a function of data lengths (i.e., one file to seven files). This result suggests that both ML training and MAP adaptation make it possible to improve the accuracy of detecting anomalies as an increase in data lengths. In this case, MAP adaptation works much better than ML training, especially when only a small amount of data is available (i.e., 6, 12, $\cdots$, 30 sec). For example, more than 83% and 73% of AUCs were reduced when using MAP adaptation instead of ML training for six and twelve seconds of adaptation data, respectively. MAP adaptation yields comparable accuracy to ML training with 42 seconds of data. This result demonstrates that MAP adaptation achieves a reliable estimation of GMM parameters at the early stage of collecting data by explicitly using the prior information obtained from other devices while ML training does not exploit any prior information, requiring larger amounts of data to develop the model.
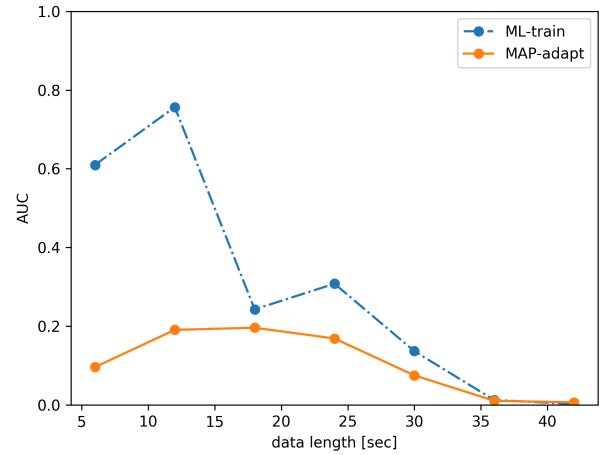

Figure 4. Effectiveness of adaptive training of normal status GMM: AUC values as a function of adaptation data size. In ML-training, adaptation data is used for training. In MAP adaptation, prior GMM is trained on training data and then adapted with adaptation data.

Table 2. Effectiveness of adaptive training of normal status GMM: AUC values as a function of length of adaptation data

data length [sec]	system	
	ML-train	MAP-adapt
6	0.60972	**0.09582**
12	0.75651	**0.19045**
18	0.24248	**0.19588**
24	0.30769	**0.16849**
30	0.13627	**0.07452**
36	0.01217	**0.01011**
42	**0.00210**	0.00624

5. CONCLUSION

In the present study, an adaptive training of a vibration-based anomaly detector for wind turbine CMS was conducted to achieve high accuracy from the early stages of monitoring. GMM was trained on data recorded from a normally operating gearbox in an experimental environment (NREL data set). For testing, run-time data observed from the gearbox in an actual wind turbine (HS data set) were input to the anomaly detector. The pre-trained GMM was adapted to the run-time data using MAP adaptation. The negative logarithmic likelihoods of run-time inputs for the adapted normal status GMM were exploited as anomaly scores. Experimental comparisons using vibration signals from different gearboxes for wind turbines (i.e., NREL and HS data sets) demonstrated that the effective use of prior information obtained from other devices in MAP adaptation yielded significant improvement in the accuracy of detecting anomalies over ML training without any prior information, especially when only small amount of data were available for parameter estimation.

6

The datasets used in the present study are relatively well organized. In the future, the effectiveness of adaptive training of a general purpose anomaly detector to the target device will be investigated on more realistic data, collected from currently available operating devices.

REFERENCES

Bartkowiak, A., & Zimroz, R. (2011). Outliers analysis and one class classification approach for planetary gearbox diagnosis. *Journal of Physics*, *305*(1).

Bechhoefer, E. (2014). *High speed gear dataset.* (http://data-acoustics.com/measurements/gear-faults/gear-1/)

Camci, F., & Chinnam, R. B. (2010). Health-state estimation and prognostics in machining processes. *IEEE Trans. on Automation Science and Engineering*, *7*(3), 581-597.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B: Methodological*, *39*, 1–22.

Deutsch, J., & He, D. (2016, Oct.). Using deep learning based approaches for bearing remaining useful life prediction. In *Proceedings of Annual Conference of the Prognostics and Health Management Society 2016.*

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Processing*, *2*(2), 291-298.

Gebraeel, N. Z., & Lawley, M. A. (2008). A neural network degradation model for computing and updating residual life distributions. *IEEE Trans. on Automation Science and Engineering*, *5*(1), 154-163.

Goodenough, D. J., Rossmann, K., & Lusted, L. B. (1974). Radiographic applications of receiver operating characteristic (ROC) curves. *Radiology*, *110*(1), 89-95.

Guclu, A., Yilboga, H., Eker, O. F., Camci, F., & Jennions, I. (2010). Prognostics with autoregressive moving average for railway turnouts. In *Proceedings of Annual Conference of the Prognostics and Health Management Society 2010.*

Hanley, A. J., & Mcneil, J. B. (1982). The meaning and use of the area under a receiver operating characteristic curve. *Radiology*, *143*(1), 29-36.

Jolliffe, I. (2002). *Principal component analysis.* Wiley Online Library.

Kim, H. ., Tan, A. C. C., Mathew, J., & Choi, B. . (2012). Bearing fault prognosis based on health state probability estimation. *Expert Systems with Applications*, *39*(5), 5200-5213.

Lee, C.-H., & Gauvain, J.-L. (1993, April). Speaker adaptation based on MAP estimation of HMM parameters. In *Proceedings of International Conference on Acoustics,* *Speech and Signal Processing* (Vol. II, p. 558-561).

Lee, C.-H., Juang, B.-H., & Lin, C.-H. (1991). A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Trans. on Signal Processing*, *39*(4), 806-814.

Liu, Z., Zuo, M. J., & Zhang, L. (2014). Remaining useful life prediction of rolling element bearings based on health state assessment. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014* (p. 86-91).

Lusted, L. B. (1971). Decision-making studies in patient management. *New England Journal of Medicine*, *284*(8), 416-424.

Medjaher, K., Tobon-Mejia, D. A., & Zerhouni, N. (2012). Remaining useful life estimation of critical components with application to bearings. *IEEE Trans. on Reliability*, *61*(2), 292-302.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in nuclear medicine*, *8*(4), 283-298.

Ogata, J., & Murakawa, M. (2016, Aug.). Vibration-based anomaly detection using FLAC features for wind turbine condition monitoring. In *Proceedings of 8th European Workshop on Structural Health Monitoring (EW-SHM)* (Vol. 3, p. 1952-1961).

Sheng, S. (2014). *Wind turbine gearbox condition monitoring vibration analysis benchmarking datasets.* (http://en.openei.org/datasets/dataset/wind-turbine-gearbox-condition-monitoring-vibration-analysis-benchmarking-datasets)

Stander, C. J. (2002). Using vibration monitoring for local fault detection on gears operating under fluctuating load conditions. *Mechanical Systems and Signal Processing*, *16*(6), 1005-1024.

Wang, T. (2012, Sept.). Bearing life prediction based on vibration signals: A case study and lessons learned. In *Proceedings of IEEE Conference on Prognostics and Health Management (PHM).*

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Ye, J., Kobayashi, T., & Higuchi, T. (2010a, Sept.). Audio-based sports highlight detection by Fourier local autocorrelations. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)* (p. 2198-2201).

Ye, J., Kobayashi, T., & Higuchi, T. (2010b, Oct.). Vowel recognition based on FLAC acoustic features and subspace classifier. In *Proceedings of International Conference on Signal Processing (ICSP)* (p. 530-533).

Ye, J., Kobayashi, T., Murakawa, M., Higuchi, T., & Otsu, N. (2012). Anomaly detection using multi-channel FLAC for supporting diagnosis of ECG. *IEEJ Trans. on Electronics, Information and Systems*, *132*(1), 111-119.

BIOGRAPHIES

Takanori Hasegawa received his B.S. degree in computer science and engineering from Waseda University, Tokyo Japan in 2017. He currently is pursuing his M.S. degree in the same university. His research interests include anomaly detection and machine learning.

Jun Ogata received the B.E., M.E., and Ph.D. degrees in electronic and information engineering from Ryukoku University, Japan, in 1998, 2000, and 2003, respectively. He is currently a senior researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interests include audio and vibration processing, spoken language understanding, and machine learning. He received the Awaya Prize Young Researcher Award from Acoustic Society of Japan (ASJ) in 2000 and Yamashita SIG Research Award from Information Processing Society of Japan (IPSJ) in 2004.

Masahiro Murakawa received his B.E., M.E., and Ph. D. degrees in mechano-informatics engineering from the University of Tokyo, Tokyo, Japan, in 1994, 1996, 1999, respectively. He heads the AI applications research team of artificial intelligence research center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. He is also a professor at the cooperative graduate school of University of Tsukuba. His research interests include real-world applications of machine learning technology, adaptive signal processing and sensor networks.

Tetsunori Kobayashi received his B.Eng., M.Eng., and Dr. Eng. degrees in electrical engineering from Waseda University, Tokyo Japan, in 1980, 1982, 1985, respectively. In 1985, he joined Hosei University as a lecturer. He had been an Associate Professor since 1987. In 1991, he moved to Waseda University. Since 1997, he has been a Professor there. He is now a Director of the Perceptual Computing Laboratory at Waseda University. He had been a Visiting Researcher at the Laboratory for Computer Science, Massachusetts Institute of Technology, at the Advanced Telecommunication Laboratory, and at the NHK Science and Technical Research Laboratory. He had been a Director of the Association for Natural Language Processing, and a chair of the IPSJ SIG on Spoken Language Processing. His research interest includes speech recognition, synthesis, enhancement, image processing, and their application to conversation robots.

Tetsuji Ogawa received his B.S., M.S., and Ph.D. in electrical engineering from Waseda University in Tokyo, Japan, in 2000, 2002, and 2005. He was a Research Associate from 2004 to 2007 and a Visiting Lecturer in 2007 at Waseda University. He was an Assistant Professor at Waseda Institute for Advanced Study from 2007 to 2012. He has been an Associate Professor at Waseda University and Egypt-Japan University of Science and Technology (E-JUST) since 2012. He was a Visiting Scholar in the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, from June to September in 2012 and from June to August in 2013, and a Visiting Scholar in Speech Processing Group, Faculty of Information Technology, Brno University of Technology, Czech Republic from June to July in 2014 and May to August in 2015. His research interests include stochastic modeling for pattern recognition, speech enhancement, and speech and speaker recognition. He is a member of the Institute for of Electrical and Electronics Engineering (IEEE), Information Processing Society of Japan (IPSJ) and Acoustic Society of Japan (ASJ). He received the Awaya Prize Young Researcher Award from the ASJ in 2011 and Yamashita SIG Research Award from the IPSJ in 2013.