Anomaly Detection Based on Feature Reconstruction from Subsampled Audio Signals

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

***Abstract*—We aim to reduce the cost of sound monitoring for maintain machinery by reducing the sampling rate, i.e., sub-Nyquist sampling. Monitoring based on sub-Nyquist sampling requires two sub-systems: a sub-system on-site for sampling machinery sounds at a low rate and a sub-system off-site for detecting anomalies from the subsampled signal. This paper proposes a feature reconstruction method for enabling anomaly detection from the subsampled signal. The method applies a long short-term memory-(LSTM)-based network for reconstructing features. The novelty of the proposed network is that it receives the subsampled time-domain signal as input directly and reconstructs the feature vector of the original signal. Experimental results indicate that our method is suitable for anomaly detection from the subsampled signal.**

***Index Terms*—machine condition monitoring, sub-Nyquist sampling, neural network, long short-term memory (LSTM)**

I. INTRODUCTION

Low-cost sound monitoring is required for maintaining machinery. Typically, skilled maintenance technicians hear the sounds from machinery and determine the overall condition. However, a shortage of skilled workers has become a serious issue, making an automated system for continuous monitoring of machinery sounds a necessity. For continuous monitoring, sensing devices must be set on-site at all times, and any recorded sounds must always be sent off-site, so the sensing cost is the most serious issue. The sensing cost consists of that of power consumption, sensing devices including analog-digital-converters (ADCs), network communication, etc. In general, these costs decrease as the data size, which corresponds to the sampling rate, decreases. Therefore, we aim to reduce the cost of monitoring by reducing the sampling rate, i.e., sub-Nyquist sampling.

We herein propose a method for subsampling the original sound and for detecting anomalies from the subsampled signal. A monitoring system based on sub-Nyquist sampling requires two sub-systems: a system on-site for sampling machinery sounds at a low rate and a sub-system off-site for detecting anomalies from the subsampled signal. The former sub-system requires clarifying what subsampling method provides sufficient performance for anomaly detection. The latter requires clarifying how can we detect anomalies from the subsampled signal. In our previous work [1], we have shown subsampling methods providing sufficient performance, i.e. random sampling, co-prime sampling, and sparse ruler sampling, and so we focus on the anomaly detection method for the latter sub-system in this paper.

The anomaly detection part of the proposed method is based on a long short-term memory-(LSTM)-based feature reconstruction. Research on anomaly detection has been done for numerous applications, e.g., acoustic surveillance [2] [3], combustion instability prediction [4], and structural health monitoring [5]. Anomaly detection is defined as detecting an outlier from a model of normality, and it can be performed by training the model of normality from a training dataset. The training dataset regularly includes only normal data, i.e., its task is unsupervised machine learning. In the field of acoustic anomaly detection, Gaussian mixture models (GMM) and hidden Markov models (HMM) have been the most widely used [2] [3]. Several works have proposed support vector machines (SVM) [4]. Also, research on anomaly detection based on neural networks has been done for many years [5][6]. In most cases, the network layout for modeling the normal behavior is an "autoencoder." Versions of the autoencoders have been improved using LSTM recurrent neural networks (RNNs), which can model a time series with a temporal correlation [7][8][9][10]. However, the conventional methods do not work for anomaly detection from the subsampled signal. To solve the problem, we propose an end-to-end approach, i.e., the proposed network receives the subsampled time-domain signal as input directly and reconstructs the feature vector yielded from the Mel spectrogram of the original signal, although conventional networks [8][9] receive the feature vector yielded from the Mel spectrogram. Hereafter, we call the conventional networks "feature-input and feature-output" (FIFO) and call the proposed network "time-domain-input and feature-output" (TIFO).

II. RELATION TO PRIOR WORK

The main contribution of this paper is to propose a feature reconstruction method for enabling anomaly detection from the subsampled signal, and the detection performance based on the proposed TIFO-type feature reconstruction is higher than that of the conventional methods such as the FIFO.

The conventional FIFO-type networks [8][9] cannot reconstruct the feature vector of the original signal. In the FIFO, the input feature vector yielded from the subsampled signal has already lost information about frequencies higher than the Nyquist frequency, and feature reconstruction does not work,

ISBN 978-90-827970-1-5

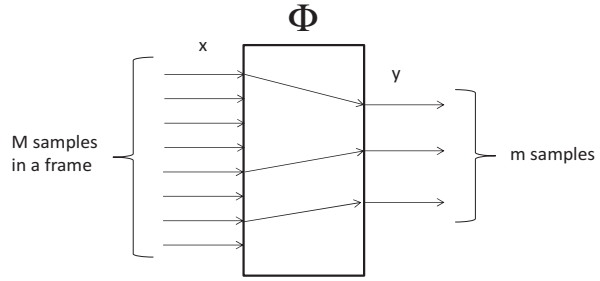


Fig. 1. Subsampling observation model

so anomaly detection also fails. In contrast, the subsampled time-domain signal keeps some information about frequencies higher than the Nyquist frequency, so the TIFO can extract it.

We have already proposed another conventional LSTM-autoencoder network [1]. This conventional network receives the subsampled time-domain signal as input and reconstructs the time-domain signal without missing data, and the network can be called "time-domain-input and time-domain-output" (TITO). The TITO-type network can also exploit information about frequencies higher than the Nyquist frequency. However, the detection accuracy of the TIFO is higher than that of the TITO because it is easier to reconstruct the feature vector than to reconstruct the original time-domain signal. That is, if the original time-domain signal is completely known, the feature vector can be calculated from it, but the reverse is not true. In Section V, we present an experimental comparison of the TIFO and the TITO.

III. PROBLEM STATEMENT

A. Anomaly Detection from Subsampled Signals

We define the real-valued discrete-time original sound of a short time frame represented by a column vector $\boldsymbol{x}$ of dimensions $M \times 1$. The observed signal $\boldsymbol{y}$ is obtained by subsampling the original sound $\boldsymbol{x}$. The $\boldsymbol{y}$ is the $m \times 1$ column vector, the elements of which are the result of the inner products between $(\boldsymbol{\Phi}_j)_{j=1}^m$ and $\boldsymbol{x}$. The $(\boldsymbol{\Phi}_j)_{j=1}^m$ corresponds to the j-th sample. Only one element of $\boldsymbol{\Phi}_j$ is 1, and the other elements are 0, and $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1|\boldsymbol{\Phi}_2|...|\boldsymbol{\Phi}_m]^T$ contains at most one nonzero entry $(= 1)$ in each row or column. Figure 1 illustrates this subsampling observation model. The following undetermined linear equation can be introduced:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}. \tag{1}$$

The problem to be solved is to detect anomalies in an unknown $\boldsymbol{x}$ from given $\boldsymbol{\Phi}$ and $\boldsymbol{y}$. In general, $\boldsymbol{\Phi}$ changes every frame, the number of candidates of $\boldsymbol{\Phi}$ may be enormous, and implementation is difficult. In this study, the sampling pattern repeats with period M to limit the number of candidates of $\boldsymbol{\Phi}$. Also, the division number D as a positive integer is introduced, i.e., the frame shift is M/D. Therefore, the number of candidates of $\boldsymbol{\Phi}$ is limited to D, and $\boldsymbol{\Phi}$ can be repeated with period D frames. This paper clarifies how we can detect anomalies from $\boldsymbol{\Phi}$ and $\boldsymbol{y}$.

B. Subsampling methods

This section explains the candidates of the sampling methods. As described in Section III-A, the sampling pattern repeats with period M, so we assume that all the methods cut the sampling pattern at the end of every frame.

1) Uniform Sampling: Uniform sampling is the simplest method. The signal is sampled at time kT using uniform sampling, where T is a unit-time interval, and

$$k \in \{0, u, 2u, 3u, \cdots\}, \tag{2}$$

where u is a constant positive integer. The ρ is defined as the number of the observed samples divided by the number of original samples. The ρ of the aforementioned case is $1/u$.

2) Random Sampling: The signal is sampled at kT using random sampling, where k is selected from all the integers using Bernoulli sampling with a probability equal to ρ. The original signal can be reconstructed at high accuracy from the signal recorded using random sampling on the condition that it is sparse in the frequency domain [11][12]. As described in Section III-A, the number of candidates of $\boldsymbol{\Phi}$ needs to be limited. Therefore, the sampling pattern randomly generated once is reused with period M.

3) Co-Prime Sampling: The signal is sampled at kT using co-prime sampling, where

$$k \in \{0, u, 2u, 3u, \cdots\} \cup \{v, 2v, 3v, \cdots\}, \tag{3}$$

and u and v are co-primes. The ρ is $(u + v - 1)/uv$.

4) Sparse Ruler Sampling: The signal is sampled at kT using sparse ruler sampling, where k is a set called a circular sparse ruler $\mathcal{K}$ such that for every $l = 0, \cdots, L-1$ there must exist at least one pair of elements $k_1, k_2 \in \mathcal{K}$ satisfying $(k_1 - k_2) \bmod L = l$ [13]. The L is the period of sparse ruler sampling. If $\mathcal{K}$ contains a minimum number of elements, it is called a minimal sparse ruler. The ρ is $|\mathcal{K}|/L$. Minimal sparse rulers have been found for certain L values, so we can choose the known minimal sparse rulers as the sampling pattern.

IV. PROPOSED FEATURE RECONSTRUCTION AND ANOMALY DETECTION

In this section, the feature reconstruction method for anomaly detection is explained. Figure 2 shows the network layout for feature reconstruction. The network consists of the input layer, one feedforward layer, two LSTM layers, and an output feedforward layer. All the layers are fully connected. The input layer has M units receiving the subsampled signal, and zeros are padded at the $(M - m)$ removed points. As described in Section I, its key point is that the layout is end-to-end, i.e., the input is the subsampled time-domain signal, and the output is the reconstructed feature vector. This end-to-end layout called the TIFO is utilized to solve the problem of the conventional FIFO-type networks not working for anomaly detection from the subsampled signal.

The output of the reconstruction network is corresponded to the feature vector yielded from the Mel spectrogram. The output feature vector is similar to that of the conventional

ISBN 978-90-827970-1-5 © EURASIP 2018

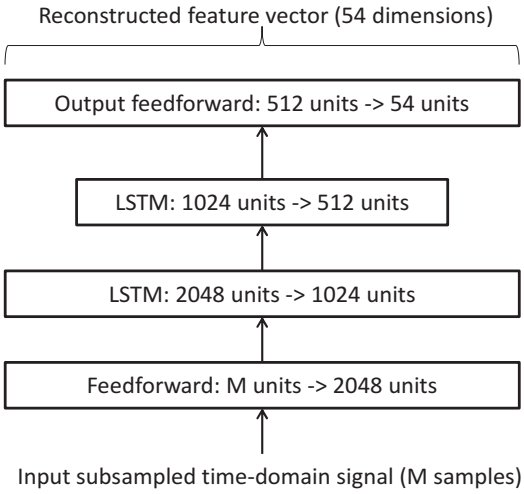


Fig. 2. Layout of feature reconstruction network

FIFO-type networks [8][9]. The STFT-frequency scale is converted to the Mel-frequency scale using a filter-bank with 26 triangular filters. The first 26 dimensions of the feature vector are the logarithmic powers of the Mel spectrogram:

$$P_{\log}(\tau, n) = \log(P(\tau, n) + 1.0), \tag{4}$$

where τ is the frame index, $n = 1, \cdots, 26$ is the Mel-frequency index, and $P(\tau, n)$ is the power of the Mel spectrogram. The next 26 dimensions are the positive first order differences:

$$D(\tau, n) = P_{\log}(\tau, n) - P_{\log}(\tau - 1, n). \tag{5}$$

Also, the frame energy and its derivative are included. Therefore, the dimension of the feature vector is 54.

Training was performed by minimizing the average error between the feature vector calculated from the original time-domain signal and the output vector over a training set. The parameters of the network were trained so that the network outputs the feature vector of the signal without missing data. The rectified linear unit (ReLU) [14] was utilized as the activation function after the feedforward layers. In addition, the hyperbolic tangent was applied as the activation function after the LSTM layers. We utilized batch normalization [15] for each layer. Batch normalization is known to be effective for training acceleration. However, in regression like this task, the scales of the input tend to be very different between frames, and the batch normalization does not work. The subsampled signal was frame-wise normalized before input to solve this problem. We applied Adam [16] for training. If the norm of a gradient was greater than one, the gradient by its norm was divided [17]. We utilized the dropout [18] for the connections before both LSTM layers, and the dropout rate was set to 0.1.

After feature reconstruction, another process for anomaly detection must be performed. Anomaly detection is an unsupervised learning task, so GMM, one-class SVM, etc. can be used as the detection process. Also, the FIFO-type autoencoders [8][9] can be applied after the proposed TIFO. In Section V, we present experimental results when using GMM.

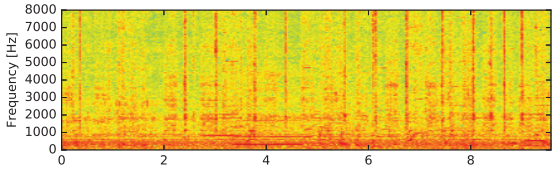


Fig. 3. Spectrogram of a part of the original sound. The X-axis shows the time in seconds.

TABLE I
CO-PRIME PAIRS FOR EXPERIMENTAL EVALUATION

u	v	ρ
5	7	0.31
7	8	0.25

V. EXPERIMENTAL RESULTS

The experimental evaluation was conducted to investigate whether or not the proposed algorithm for anomaly detection can work well. The signals subsampled using different sampling methods were used as the observed signal, different reconstruction algorithms were applied for the subsampled signals, and the results of anomaly detection were compared.

To show industrial usefulness, we used the sound of a real automated machine. The machine has a lot of parts such as mechanical arms and continues to do a same task repeatedly. When a machine repeatedly performed a series of work, the original sound was recorded (See Fig. 3). The total length of the original sounds was 10 minutes of 16-bit audio signals sampled at 16 kHz. The sounds were divided into 5 minutes of data for training and 5 minutes of data for evaluation, with both of them corresponding to the normal condition. Because each of the parts works multiple times within several seconds, the training data of 5 minutes includes a sufficient amount of the sounds from each of the parts. To generate abnormal evaluation data, we randomly selected a frequency for each test from 4, 5, 6, and 7 kHz, and the sinusoidal wave of the selected frequency was mixed into the normal evaluation data, where the power ratio between the sinusoidal wave and the original sound was set to -15 dB. Actually, in many cases, anomalies cause friction, and it tends to amplify certain frequencies, so anomalies were simulated by adding sinusoidal waves.

The Hanning window with the frame size of $M = 1024$ and the frame shift of 512 were applied, so the division number was $D = 2$. The length of the sequence fed to the network was 16 frames, and anomaly detection was used only for these frames. The mini-batch size was set to 256. The co-prime pairs shown in Table I were chosen for co-prime sampling. The minimal circular sparse rulers shown in Table II were chosen for sparse ruler sampling. These minimal circular sparse rulers were generated from length-$\lfloor L/2 \rfloor$ minimal linear

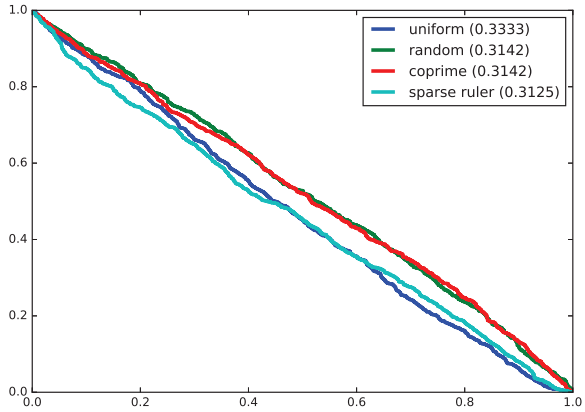
ISBN 978-90-827970-1-5

TABLE II
CIRCULAR SPARSE RULERS $\mathcal{K}$ FOR EXPERIMENTAL EVALUATION

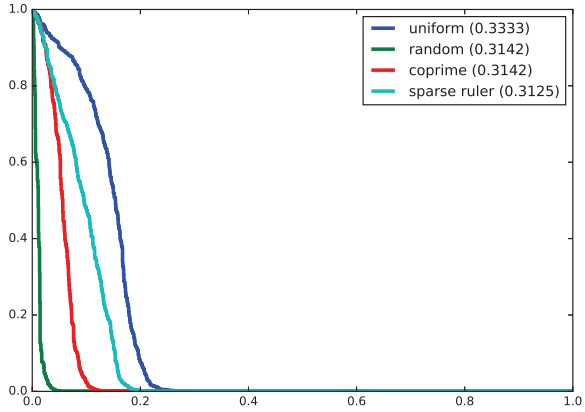
$L-1$	$\lfloor L/2 \rfloor$	$\mathcal{K}$	$\lvert\mathcal{K}\rvert$	ρ
15	8	$\{0,1,5,6,8\}$	5	0.31
23	12	$\{0,1,7,8,10,12\}$	6	0.25

sparse rulers. The following three reconstruction methods were utilized: (a) The first was an FIFO-type network similar to [8][9], the input and the output of which are the 54-dimensional feature vector of the subsampled signal and that of the original sound, respectively. For the input layer, the number of units was changed to 54. The hidden layers were the same as those in Fig 2. (b) The second was the conventional TITO-type network [1], the input and the output of which are the time-domain signal of the subsampled signal and that of the original sound, respectively. The number of units was changed to the frame size M for both the input layer and the output layer. The hidden layers were the same as those in Fig 2. (c) The third was the proposed TIFO-type network, the layout of which is shown in Fig 2. For (a) and (c), GMM was applied in anomaly detection after feature reconstruction, and the number of mixture components was 8. For (b), identically to the previous work [1], anomalies were detected by thresholding the average error at the sampling points between the subsampled signal and the reconstructed signal.

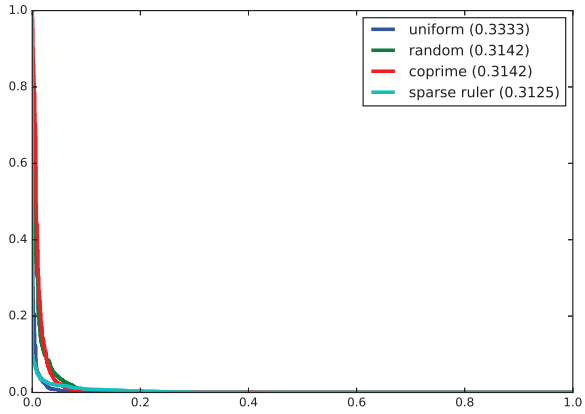
Figures 4 and 5 show the evaluation results. These results show that anomaly detection based on the FIFO failed and that the other two approaches succeeded. They indicate that difficulty occurs in reconstructing the feature vector from that of the subsampled signal; also, the FIFO-type reconstruction is not suitable for anomaly detection from the subsampled signal. Furthermore, a comparison of (b) with (c) shows that the detection performance of the proposed method was higher than that of the TITO. These results indicate that the proposed TIFO-type network improves the detection performance as described in Section II. In addition, the results of (b) were that the detection performance of uniform sampling was far lower than that of non-uniform sampling methods, whereas the results of (c) were that the detection performance of uniform sampling was at the same sufficient level as that of non-uniform sampling methods. Considering that aliasing must occur in the case of uniform sampling, these results indicate that the proposed TIFO can extract information about frequencies higher than the Nyquist frequency even from aliasing components mixed with low-frequency components.



(a) FIFO



(b) TITO [1]



(c) Proposed TIFO

Fig. 4. ROC curves in the case that ρ is about 0.33. X and Y show the false positive rate and the false negative rate respectively.

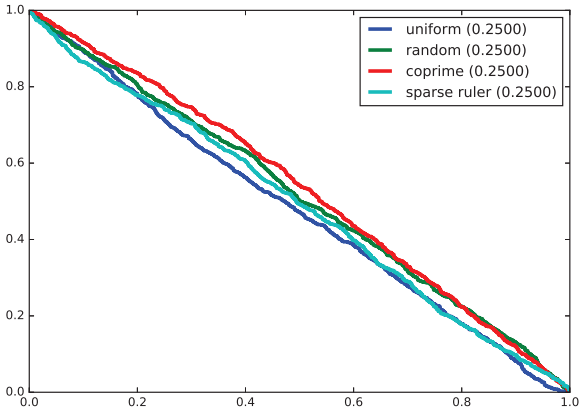
VI. CONCLUSION

To reduce the cost of sound monitoring, we proposed a feature reconstruction method for anomaly detection from the subsampled audio signals. The proposed method is based on the LSTM-based network called TIFO, and its key point is that it receives the subsampled time-domain signal and reconstructs the feature vector of the original signal. The TIFO-type structure enables detecting anomalies from the subsampled signal. Experimental results showed that the proposed TIFO is suitable for anomaly detection from the subsampled signal.

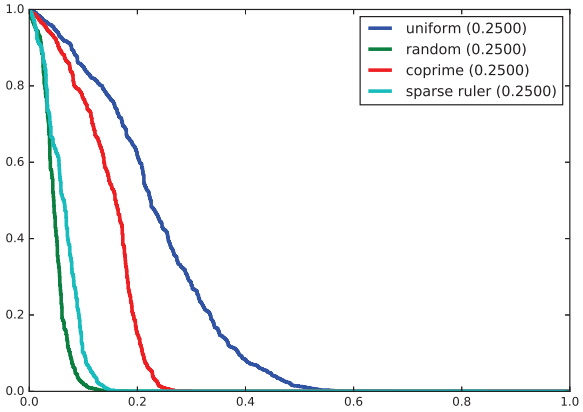
REFERENCES

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

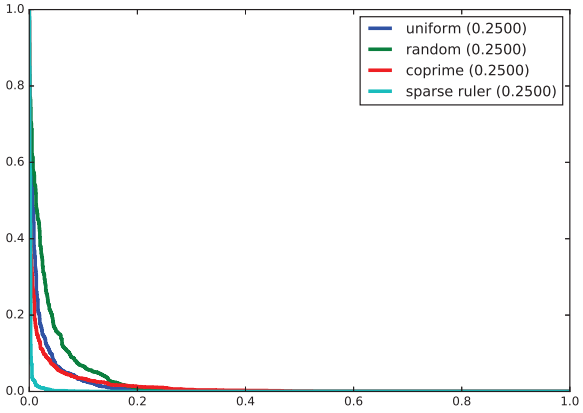
ISBN 978-90-827970-1-5



(a) FIFO



(b) TITO [1]



(c) Proposed TIFO

Fig. 5. ROC curves in the case that ρ is about 0.25. X and Y show the false positive rate and the false negative rate respectively.

[2] 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.

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

[4] L.A. Clifton, H. Yin, and Y. Zhang, "Support vector machine in novelty detection for multi-channel combustion data," in *Proc. 3rd International Symposium on Neural Networks*, May 2006, pp. 836–843.

[5] K. Worden, G. Manson, and D. Allman, "Experimental validation of a structural health monitoring methodology: Part I. Novelty detection on a laboratory structure," *Journal of Sound and Vibration*, vol. 259, no. 2, pp. 323–343, Jan. 2003.

[6] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proc. 2nd Workshop on Machine Learning for Sensory Data Analysis (MLSDA)*, Dec. 2014, pp. 4:4–4:11.

[7] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc. 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Apr. 2015, pp. 89–94.

[8] 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. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 1996–2000.

[9] 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. 2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7.

[10] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo, "Unsupervised extraction of video highlights via robust recurrent auto-encoders," in *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 4633–4641.

[11] J. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, A. Gilbert, M. Iwen, and M. Strauss, "Random sampling for analog-to-information conversion of wideband signals," in *Proc. IEEE Dallas/CAS Workshop on Design, Applications, Integration and Software (DCAS)*, Oct. 2006, pp. 119–122.

[12] P. Maechler, N. Felber, and A. Burg, "Random sampling adc for sparse spectrum sensing," in *Proc. The 19th European Signal Processing Conference (EUSIPCO)*, Aug. 2011, pp. 1200–1204.

[13] D. Romero, D.D. Arianandа, Z. Tian, and G. Leus, "Compressive covariance sensing: Structure-based compressive sensing beyond sparsity," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 78–93, Jan. 2016.

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

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

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

[17] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," in *arXiv:1211.5063*, 2012.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

ISBN 978-90-827970-1-5