International Journal of Advanced Smart Convergence Vol.9 No.2 20-27 (2020)
http://dx.doi.org/10.7236/IJASC.2020.9.2.20

IJASC 20-2-3

CNN based Sound Event Detection Method using NMF Preprocessing in Background Noise Environment

Bumsuk Jang[1], Sang-Hyun Lee[2†]

[1]*CEO, BS SOFT Co., LTD., Gwangju, Korea*
[2,†]*Assistant Professor, Department of Computer Engineering, Honam University, Gwangju, Korea*

Abstract

Sound event detection in real-world environments suffers from the interference of non-stationary and time-varying noise. This paper presents an adaptive noise reduction method for sound event detection based on non-negative matrix factorization (NMF). In this paper, we proposed a deep learning model that integrates Convolution Neural Network (CNN) with Non-Negative Matrix Factorization (NMF). To improve the separation quality of the NMF, it includes noise update technique that learns and adapts the characteristics of the current noise in real time. The noise update technique analyzes the sparsity and activity of the noise bias at the present time and decides the update training based on the noise candidate group obtained every frame in the previous noise reduction stage. Noise bias ranks selected as candidates for update training are updated in real time with discrimination NMF training. This NMF was applied to CNN and Hidden Markov Model(HMM) to achieve improvement for performance of sound event detection. Since CNN has a more obvious performance improvement effect, it can be widely used in sound source based CNN algorithm.

Key words: *Non-negative matrix, CNN, artificial neural networks, Sound Event Detection, Signal to Noise Ratio.*

1. INTRODUCTION

Sound events such as screams, gunshots, glass breaks, and so on, are often associated with critical or noteworthy situations. The automatic detection and monitoring of these sound events can be of great use for surveillance purposes. Compared to traditional surveillance systems based on video cameras, audio sensors are insensitive to illumination or occlusion, cheaper, and more suitable for privacy. Moreover, some events like gunshots have no evident visual characteristics and are more suited to audio detection. Nowadays, because of these advantages, the audio information has been exploited solely or jointly with video signals in intelligent surveillance systems.

The primary objective of a Sound Event Detection (SED) system is to identify the type of sound source present in an audio clip or recording and returns the onset and offset of the identified source. Such a system has great potential in several domains such as activity monitoring, environmental context understanding, and

Manuscript Received: March. 17, 2020 / Revised: March. 21, 2020 / Accepted: March. 25, 2020
Corresponding Author: leesang64@honam.ac.kr
Tel: +82-62-940-5285, Fax: +82-62-940-5285
Assistant Professor, Department of Computer Engineering, Honam University, Gwangju, Korea

multimedia event detection [1], [2]. Generally, research on sound event detection methods has been focused on extracting discriminating audio features, and training effective classifiers for distinguishing different sound classes and noise (see [3] and [4] for a complete review). The input audio signal is typically transformed to the time-frequency domain, and is represented by features like the mel-scale spectral energies, or simply, the magnitude spectrogram. Commonly used classifiers include Gaussian mixture models (GMMs), support vector machines (SVMs), artificial neural networks (ANNs), and non-negative matrix factorization (NMF). More recently, deep learning methods, which are receiving increasing interest, have also been studied for sound event detection, if enough training data are available [5]. However, there are several challenges associated with SED in real life scenarios.

Firstly, in real-life scenarios, different sound event can occur simultaneously [2]. Secondly, the presence of background noise could complicate the identification of sound event within a particular time frame [6]. This problem is further aggravated when the noise is the prominent sound source resulting in a low Signal to Noise Ratio (SNR).

Thirdly, each event class is made up of different sound sources, e.g. a dog bark sound event can be produced from several breeds of dogs with different acoustic characteristics [1]. Most of the existing methods employing a well-trained classifier over the noise training set cannot handle unseen noise, and also lack the adaptation ability to time-varying noise. Performance of these methods can be severely reduced when the test condition does not match that of the training data, caused by either different recording devices or locations. How to reduce noise, and more importantly, how to generalize well to unknown or changing noise conditions remains a great challenge for sound event detection methods, and is also the focus of this work [8].

Although a large number of different SED system were proposed in the past, a majority of them were mainly based on Gaussian Mixture Model (GMM) [11], Hidden Markov Model (HMM) [10] or the use of dictionaries constructed using NMF [12-14]. However, due to the rising success of deep learning in other domains [15-18], deep learning for SED development is now a norm and has been shown to perform slightly better than established methods [1]. In this paper, we propose an adaptive noise reduction method for sound event detection based on NMF. We demonstrated the performance improvement of SED by applying NMF to HMM and CNN. Experimental results show that it is appropriate to apply NMF to SED using CNN.

2. RELATED WORKS

2.1 NMF

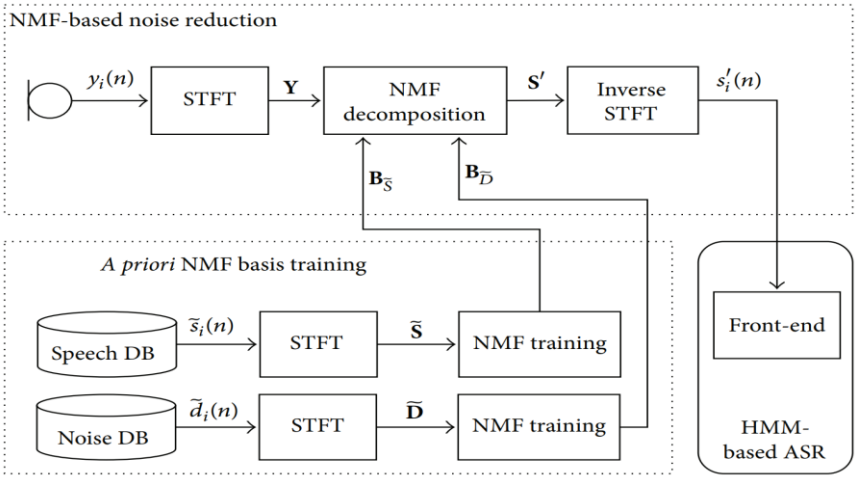

Figure 1. Procedure of the conventional NMF-based noise reduction method.

The NMF popularized by Lee and Seung [21] is an effective method to decompose a non-negative matrix, $M \in \mathbb{R}^{\geq 0, L \times N}$, into two non-negative matrices, $W \in \mathbb{R}^{\geq 0, L \times R}$ and $H \in \mathbb{R}^{\geq 0, RL \times N}$. Where R is the number of components. Therefore, it can be represented as

$$M \approx WH \tag{1}$$

Where W can be interpreted as the dictionary matrix and H can be interpreted as the activation matrix. These two matrices can be randomly initialized and updated through the multiplicative rule given as [23]

$$W \leftarrow W \otimes \frac{W^T \frac{M}{WH}}{W^T 1} \tag{2}$$

$$H \leftarrow H \otimes \frac{\frac{M}{WH} H^T}{1 H^T} \tag{3}$$

W is commonly extracted on isolated events to form a dictionary and SED is performed by applying a threshold on the activation matrix obtained from the decomposition of the test data [11]. Since NMF only works on non-negative matrix, it was applied on the Mel spectrogram prior to the logarithm operation. Thus, M represent the Mel spectrogram with L as the number of Mel bins and N as the number of frames. In this paper, instead of consolidating W to form the dictionary. We find the H to indicate which frames of each audio clip are activated (above a pre-defined threshold) to label the weakly labelled data so that the weakly labelled data becomes an approximated strongly labelled data.

In this system, training inputs are Mel-frequency scaled. This is because they can provide a reasonably good representation of signal's spectral properties. At the same time, they also provide reasonably high inter-class variability to allow class discrimination by many different machine learning approaches [19].

2.2 Sound Event Detection

In the recent years, SED development has been overwhelmed with the use of deep learning algorithms particularly the use of CNN or Convolutional Recurrent Neural Network (CRNN). This phenomenon was also reflected in the 2018 DCASE challenge, where almost all participants for Task 4 (Large-scale weakly labeled semi-supervised sound event detection in domestic environments) proposed the use of CRNN [9]. As discussed in [1], CNN has the benefit of learning filters that are shifted in both time and frequency while Recurrent Neural Network (RNN) has a benefit of integrating information from the earlier time windows. Thus, a combined architecture has the potential to benefit from two different approaches that suggest its popularity.

The CRNN architecture proposed by Cakir et al. [1] first extracted features through multiple convolutional layers (with small filters spanning both time and frequency) and pooling in the frequency domain. The features were then fed to recurrent layers, whose features were used to obtain event activity probabilities through a feedforward fully connected layer. Evaluation over four different datasets had also shown that such a method has a better performance as compared to CNN, RNN and other established SED system. However, such a system would require a large amount of annotated data for training.

Lu [10] proposed the use of Mean Teacher Convolution System that won the DCASE 2018 Task 4 challenge with an F1 score of 32.4%. In their system, context gating was used to emphasize the important parts of audio features in frames axis. Mean-Teacher semi-supervised method was then applied to exploit the availability of unlabeled data to average the model weights over training steps. Although, this system won the 2018 challenge, there is still a large room for improvement.

3. PROPOSED METHOD

3.1 Audio Processing

In this system, training inputs are mel-frequency scaled. This is because they can provide a reasonably good representation of signal's spectral properties. At the same time, they also provide reasonably high inter-class variability to allow class discrimination by many different machine learning approaches [19].

In this paper, audio clips were first resampled to 16 kHz that were suggested to contain the most energies [20]. Moreover, segments containing higher frequency may not be useful for event detection in daily life [10]. A short-time fast Fourier transform with a Hanning window size of 1024 samples and a hop size of 500 samples was used to tabulate the spectrogram. After that, a mel filter bank of 64 and bandpass filter of 50 Hz to 14 kHz was applied to obtain the mel spectrogram to be used as input to the training model. Finally, a logarithm operation was applied to obtain the log mel spectrogram.

3.2 Adaptive NMF

In this section, an NMF-based adaptive noise sensing and reduction method is proposed to mitigate the degradation of noise reduction when there is a mismatch in noise types between noise basis training and estimation using NMF. Figure 2 shows the procedure of the proposed NMF-based adaptive noise sensing and reduction method. As shown in the figure, the procedure is divided into three different processing stages: a priori NMF basis modeling, NMF-based adaptive noise sensing, and noise reduction.

The first processing stage of the proposed method is the same as that of the conventional method described in Section 2. In other words, clean speech signals and noise signals are separately applied to the NMF training in order to obtain the a priori basis matrices. In the second processing stage, the adaptive noise sensing is performed to decompose the noisy input spectrum into speech and noise spectrum using a priori speech basis matrix estimated by the first processing stage. That is, the noise basis and activation matrices are obtained by adapting a priori noise basis from the instantaneous noise frames of the noisy input signal. Finally, the third processing stage of the proposed method estimates the noise-reduced speech signal by constructing a Wiener filter [24] using the adaptively estimated noise spectrum.

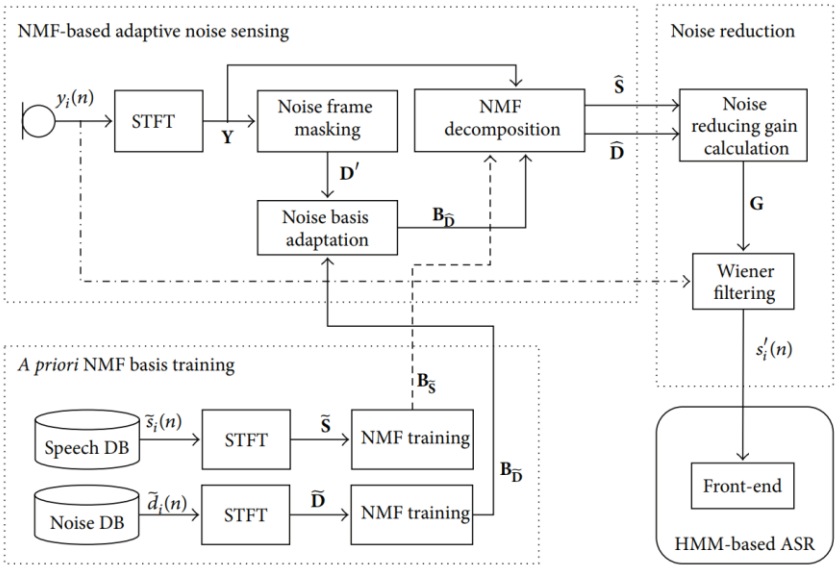

Figure 2. Procedure of the proposed NMF-based adaptive noise sensing and reduction method.

3.3 Convolutional Neural network

The CNN used in this system is modified based on the one proposed in Kong et al. [22] proposed four different CNN with a different number of layers and pooling operators and found that the nine layers CNN with max pooling operator achieved the best performance [7]. In this paper, we are interested in finding out whether with the inclusion of NMF, will a shallower CNN produce a comparable or even a better result.

As shown in Fig 1., a 3 layers CNN with 1 layer Fully Connected Network with Softmax is proposed. In this architecture, it consists of 3 convolutional layers of kernel size 5 x 5 with a padding size of 2 x 2 and strides 1 x 1. This architecture is almost similar to Mesaros et al. [21] except for the kernel size and the number of layers.

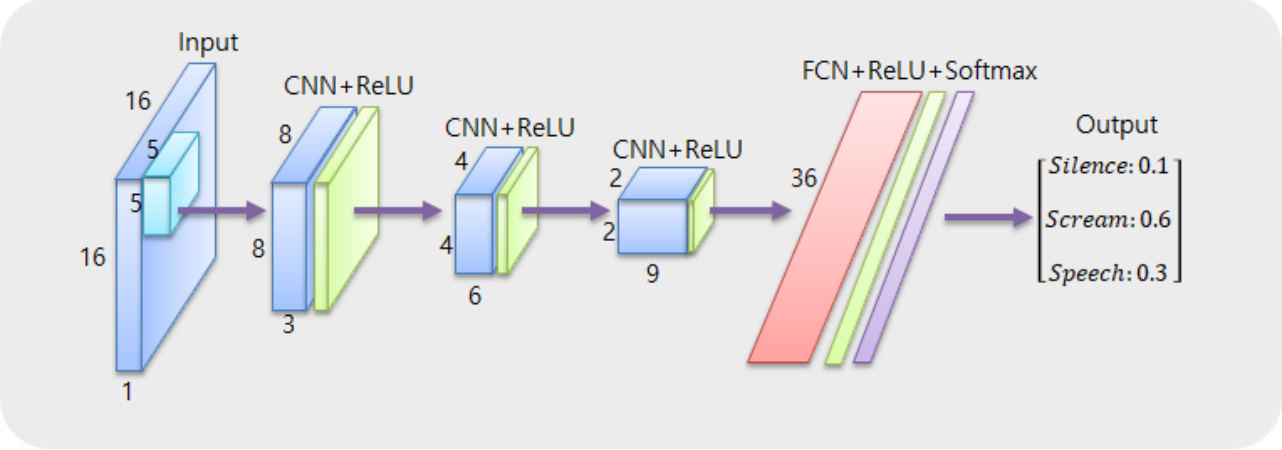

Figure 3. Architectures of Convolution Neural Network

3.4 System Flow

This paper proposes a supervised and adaptive NMF framework for sound event detection, as shown in Fig 1. The input audio signals are first processed via the short-time Fourier transform (STFT), and magnitude spectrograms are used for audio signal representation. The detection method has two phases—a training phase and a test phase. During training, for each sound event class, an event dictionary is learned using its clean event training data. The spectrograms of all of the event training signals for a specific class are concatenated to yield a data matrix denoted by V_s^train∈R_+^(F×Ttrain), and the standard NMF is then performed according to Equations (2) and (3). The resulting event dictionary Ws is used and kept fixed during the test. In the test phase, the input noisy signal is processed following the three steps of noise dictionary learning, source separation, and event detection. It should be mentioned that the present algorithm is developed in an offline manner, and real-time processing is not emphasized in this paper. For an input test signal, a noise dictionary is estimated from the current input, and then used in the supervised separation process combined with the pre-trained event dictionary. Meanwhile, the time-frequency weights for A-NMF are derived according to prior information of the target event class, as well as from the results of the noise estimation. After source separation, the event spectrogram is reconstructed and post-processed by an energy detector so as to generate the detection results.

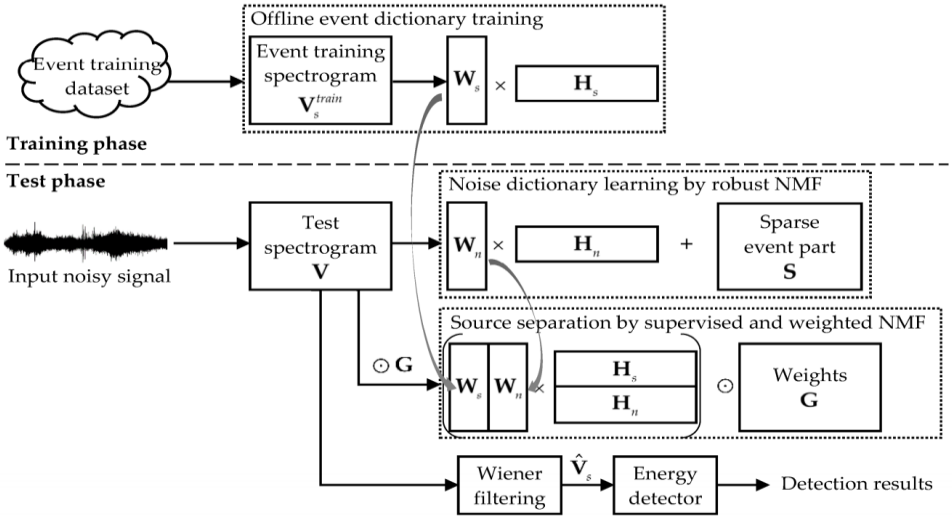

Figure 4. Framework of the proposed sound event detection method based on non-negative matrix factorization (NMF)

4. EXPERIMENT

For the experiment, assuming the actual situation, we experimented with directors and actors to detect the screams that occur with various noise environments indoor space with reverberation size of 33m^2. The training data used noise and scream dataset of DCASE(2017) which is most famous worldwide Challenge of SED. Based on the training data, we implemented the adaptive NMF and evaluated the noise canceling performance and SED performance of Scream based on the field data collected assuming the actual environment. The figure 5 shows the spectrogram comparisons of the attenuated noise from the collected field data using Adaptive NMF. The figure 5 (a), (b) shows that the noise is eliminated, but the features for identifying the main sound source event remain.

As shown in Fig. 5 (c) and (d), the performance of HMM and CNN-based SED has been improved since preprocessed with A-NMF has reduced the False Alarm which is detected as a target(scream) sound but is not. The effect was greater with CNN (fig 5. (d)) than with HMM (fig 5. (c)). When using CNN, the threshold for event discrimination threshold was 1.2, and after A-NMF was applied, the threshold value was less sensitivity than before. Thus, detection performance is more stable than without A-NMF. In the CNN-based SED using A-NMF, we addressed that it is very accurate in determining the sound source event of real noise environment.

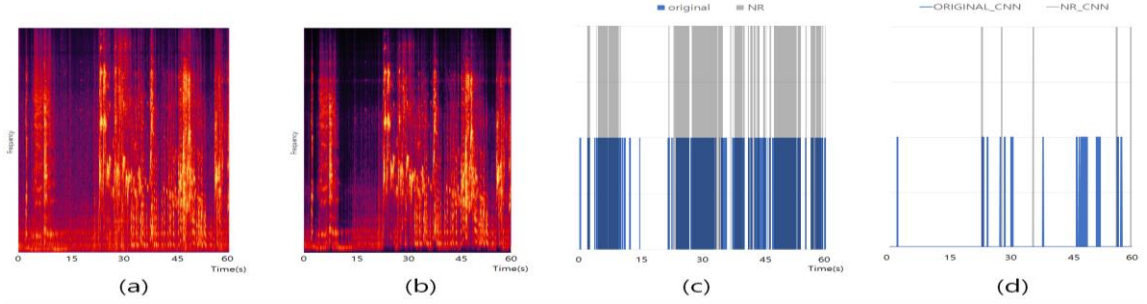

Figure 5. Spectrogram for noise reduction: left(a) and right columns(b) show spectrograms of original sound data and noise reduced sound data, respectively. And Comparison of SED performance between original data and noise reduced data: using HMM (c) and CNN (d)

5. CONCLUSION

In this paper, an adaptive noise reduction method based on supervised and adapted NMF is proposed for sound event detection in non-stationary background noise environment. The proposed adaptive strategies are guided by both the prior knowledge of sound events and the results from noise estimation, which provide an additional discriminating ability to the original NMF model. For one thing, the weight of each frequency band is quantified as a trade-off between its contributions to constructing the target event class and noise. This forces the NMF decomposition to emphasize those distinct or dominant frequencies of the target event class more. The frequency weighting scheme has shown its effectiveness in improving discrimination when dealing with strong interfering sounds with highly overlapping frequency components. Of all of the adaptive schemes, the experimental results show that the best performance is achieved by the combined time-frequency scheme that makes the best use of prior knowledge.

As the proposed method employs a noise estimation technique from the current input noisy signal, which also guides the derivation of both the frequency and statistical noise biases, the system can be easily adapted to different and time-varying noise conditions. Nevertheless, to ensure performance, the sound events in the training set and the development/evaluation set should better come from the same distribution. In the present algorithm, an average spectral template is extracted for representing a sound event class when determining noise biases, which has limitations in dealing with the diversity of characteristics within a sound class. Future work will address the adaptation of the proposed approach with multiple templates or templates considering the temporal dynamics of sound events. In addition, another improvement of the present algorithm would be supporting it with real-time processing by using a sliding window, which would make this work more promising for practical use.

REFERENCES

[1] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection,” IEEE/ACM Trans Audio, Speech, an Language Process., vol. 25, no. 6, pp. 1291-1303, Jun. 2017. DOI: 10.1109/TASLP.2017.2690575

[2] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, “Duration-Controlled LSTM for Polyphonic Sound Event Detection,” IEEE/ACM Trans Audio, Speech, an Language Process., vol. 25, no. 11, pp. 2059-2070, Nov. 2017. DOI: 10.1109/TASLP.2017.2740002

[3] Crocco, M.; Cristani, M.; Trucco, A.; Murino, V. Audio surveillance: A systematic review. ACM Comput. Surv. 2016, 48, 52. DOI: 10.1145/2871183

[4] Sharan, R.V.; Moir, T.J. An overview of applications and advancements in automatic sound recognition. Neurocomputing 2016, 200, 22–34. doi.org/10.1016/j.neucom.2016.03.020

[5] Cakir, E.; Parascandolo, G.; Heittola, T.; Huttunen, H.; Virtanen, T. Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans. Audio Speech Lang. Process. 2017, 25, 1291–1303. DOI: 10.1109/TASLP.2017.2690575

[6] B. McFee, J. Salamon, and J. P. Bello, “Adaptive Pooling Operators for Weakly Labeled Sound Event Detection,” IEEE/ACM Trans Audio, Speech, an Language Process., vol. 26, no. 11, pp. 2180-2193, Apr. 2018.

[7] S. Adavanne, P. Pertila, and T. Virtanen, “Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network,” Detection and Classification of Acoustics Scenes and Events 2017, Munich, Germany, Nov. 2017, pp. 1-5. DOI: 10.1109/ICASSP.2017.7952260

[8] J. Lu, “Mean Teacher Convolution System For DCASE 2018 Task 4,” Detection and Classification of Acoustics Scenes and Events 2018, Shanghai, China, Jul. 2018, pp. 1-5.

[9] D. Su, X. Wu, L. Xu, "GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection," 2010 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP), Dallas, TX, USA, Mar. 2010, pp. 4890-4893. DOI: 10.1109/ICASSP.2010.5495122

[10] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen "Acoustic Event Detection in Real Life," 18th European Signal Process. Conf., Aalborg, Denmark, Aug. 2010, pp. 1267-1271.

[11] V. Bisot, S. Essid, and G. Richard, "Overlapping Sound Event Detection with Supervised Nonnegative Matrix Factorization," 2017 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 31-35. DOI: 10.1109/ICASSP.2017.7951792

[12] T. Komatsu, Y. Senda, and R. Kondo, "Acoustics Event Detection Based on Non-Negative Matrix Factorization With Mixtures of Local Dictionaries and Activation Aggregation," 2016 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP), Shanghai, China, Mar. 2016, pp. 2259-2263. DOI: 10.1109/ICASSP.2016.7472079

[13] Z. Md. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems," IEEE Commun. Surveys Tutorials, vol. 19, no. 4, pp. 2432-2455, 2017. DOI: 10.1109/COMST.2017.2707140

[14] Z. Liu, Z. Jia, C. Vong, S. Bu, J. Han, and X. Tang, "Capturing High-Discriminative Fault Features for Electronics-Rich Analog System via Deep Learning," IEEE Trans. Indust. Inform., vol. 13, no. 3, pp. 1213- 1226, Jun. 2017. DOI: 10.1109/TII.2017.2690940

[15] M. He and D. He, "Deep Learning Based Approach for Bearing Fault Diagnosis," IEEE Trans. Indust. Applications, vol. 53, no. 3, pp. 3057-3065, Jun. 2017. DOI: 10.1109/TIA.2017.2661250

[16] T. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?," IEEE Trans. Image Process., vol. 24, no. 12, pp. 5017-5032, Dec. 2015. DOI: 10.1109/TIP.2015.2475625

[17] A. Mesaros, T. Heittola , E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," IEEE/ACM Trans Audio, Speech, an Language Process., vol. 26, no. 2, pp. 379-393, Feb. 2018. DOI: 10.1109/TASLP.2017.2778423

[18] Q. Kong, Y. Cao, T. Iqbal, Yong Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio-tagging, sound event detection spatial localization: DCASE 2019 baseline systems," arXiv: 1904.03476, pp. 1-5.

[19] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788-791, Oct. 1999.

[20] Y. Xie, Z. Liu, Z. Yao, and B. Dai, "Improved two-stage Wiener filter for robust speaker identification," in Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06), pp. 310–313, Hong Kong, August 2006. DOI: 10.1109/ICPR.2006.696