NONNEGATIVE MATRIX FACTORIZATION-BASED SOURCE SEPARATION WITH ONLINE NOISE LEARNING FOR DETECTION OF RARE SOUND EVENTS

Kwang Myung Jeon, Nam Kyun Kim, and Hong Kook Kim[*]

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology
123 Cheomdangwagi-ro, Gwangju 61005, Korea
{kmjeon, skarbs001, hongkook}@gist.ac.kr

ABSTRACT

In this paper, a source separation method based on nonnegative matrix factorization (NMF) with online noise learning (ONL) is proposed for the robust detection of rare sound events. The proposed method models the rare sound event into combinations of acoustic dictionaries, which consist of multiple spectral bases. In addition, ONL is adopted during the separation to improve the robustness against unseen noises. The spectra of the sound event separated by the proposed method act as a feature vector for the deep neural network (DNN)-based binary classifier, which determines whether the event has occurred. The evaluation results using the DCASE 2017 Task 2 Dataset show that the proposed source separation method improved the F-score of the baseline DNN classifier by 6.30% while decreasing the error rate by 14.81% on average.

Index Terms—Sound event detection, nonnegative matrix factorization, online noise learning, unseen noise

1. INTRODUCTION

The detection of atypical events, such as an unforeseen accident or crash, has been consistently required for its importance regarding social safety [1], [2]. Most event detection systems so far have been based on visual event detection (VED) technology [1]. However, VED might fail to detect atypical events due to environmental conditions, such as lighting, obstacles, or a limited visual angle. On the other hand, sound event detection (SED) has been highlighted as complementing the limitations of the VED system [2]. Since SED can cover large fields for surveillance regardless of environmental conditions that limit the use of VED, it is highly suitable for use in applications regarding security or safety [3].

To apply SED to real-life surveillance tasks, various types of unseen noises should be carefully considered while developing the SED system. In other words, the SED system should reject any kind of sound as noise except the target sound event that has been registered in advance. To address this issue, nonnegative matrix factorization (NMF)-based source separation has been extensively utilized for SED tasks [2, 4–6]. NMF is suitable for separating the target sound events from the background noise if the basis for both the target sounds and the background noise is given in advance. For this reason, some conventional SED methods successfully detect target sound events in a noisy environment by combining both NMF-based source separation and the hidden Markov model (HMM)-based binary classifier [4, 5]. However, these methods might not be suitable for real-world applications because they do not consider various types of unseen noises, but rely on pre-trained bases for both the target sound events and the noise ones. Recently, semi-supervised NMF-based SED with noise dictionary learning achieved satisfactory results in the detection and classification of acoustic scenes and events (DCASE) 2016 task 2 [6]. Owing to the noise learning by the semi-supervised NMF, detection accuracy was substantially improved compared with the conventional NMF method at the low signal-to-noise ratio (SNR) of -6 dB. However, it was reported in our previous work that the semi-supervised NMF-based source separation could learn the noise dictionary from other sources in the mixture; thus, this substantially degraded the separation performance in the results [7]. For this reason, a more sophisticated source separation technique with consideration of unseen noises is required for the improved performance of SED systems.

To deal with unseen noises that might interfere with SED operation, this paper proposes a source separation method based on NMF with online noise learning (ONL) for the detection of rare sound events. The proposed method models the target sound event into combinations of corresponding dictionaries, which consist of multiple spectral bases. The dictionary for the target sound event is trained in advance by using K-means clustering [8] and unsupervised sparse-NMF (SNMF) [9]. In the source separation step, the proposed method first conducts supervised SNMF to separate the input signal into two classes: sound event and noises. After that, the ideal ratio mask (IRM) is estimated from the separated spectra by using minimum mean square error (MMSE) filtering [7]. The IRM is then applied to the input spectral power to obtain the spectral power of the sound event and noise, which has fewer artifacts than the separated result by SNMF. The spectral power of the sound event by the IRM is then fed into the deep neural network (DNN)-based binary classifier [10] as a feature vector. Meanwhile, ONL obtains the spectral basis from recently estimated noise spectral powers, then updates the noise dictionary, which is recursively fed into the SNMF separation for the subsequent input signal.

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Government of Korea (MSIT) (No. 2015R1A2A1A05001687), and by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2017-0-00255, Autonomous digital companion framework and application).

[*]Corresponding Author: Hong Kook Kim (hongkook@gist.ac.kr)

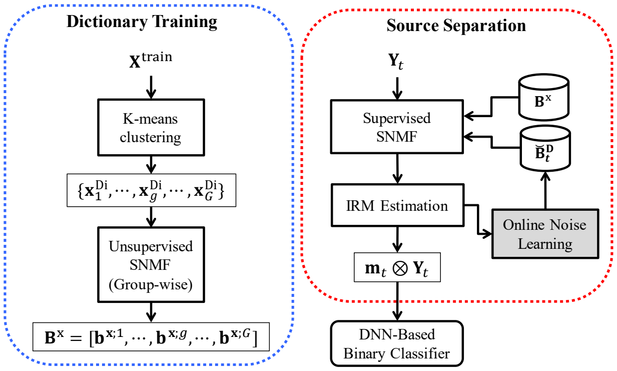

Dictionary Training
$\mathbf{X}^{\text{train}}$
K-means clustering
$\{\mathbf{x}_1^{\text{Di}}, \cdots, \mathbf{x}_g^{\text{Di}}, \cdots, \mathbf{x}_G^{\text{Di}}\}$
Unsupervised SNMF (Group-wise)
$\mathbf{B}^{\text{x}} = [\mathbf{b}^{\text{x};1}, \cdots, \mathbf{b}^{\text{x};g}, \cdots, \mathbf{b}^{\text{x};G}]$
Source Separation
$\mathbf{Y}_t$
Supervised SNMF
$\mathbf{B}^{\text{x}}$
$\breve{\mathbf{B}}_t^{\text{D}}$
IRM Estimation
Online Noise Learning
$\mathbf{m}_t \otimes \mathbf{Y}_t$
DNN-Based Binary Classifier

Figure 1: Procedure of the proposed source separation method for sound event detection.

Following this introduction, Section 2 proposes an SNMF-based source separation method with ONL. Subsequently, Section 3 evaluates the performance of the proposed method and compares it with that of a baseline DNN classifier provided by DCASE 2017 Task 2 [10]. Finally, Section 4 concludes the paper.

2. PROPOSED SOURCE SEPARATION FOR SOUND EVENT DETECTION

Figure 1 shows the procedure of the proposed source separation method-based SNMF and ONL. As indicated in the figure, the proposed method consists of the dictionary training and source separation steps. The following subsections describe each stage of the proposed method in detail.

2.1. Dictionary training

Assume that a training signal that consists of multiple sound clips of the target event is prepared. The n-th sample at the t-th frame of the training signal, $x_t^{\text{Tr}}(n)$, is transformed into the K-point mel-spectral power, $\left|X_t^{\text{Tr}}(k)\right|^2$, by using the short-time Fourier transform (STFT) and STFT-to-mel conversion matrix. Next, the P series of $X_t^{\text{Tr}}(k)$ are grouped into a spectrogram patch, $\mathbf{x}_t^{\text{Tr}}$, as

$$\mathbf{x}_t^{\text{Tr}} = \left\{\left|X_{t-P+1}^{\text{Tr}}(k)\right|^2, \cdots, \left|X_t^{\text{Tr}}(k)\right|^2\right\}. \tag{1}$$

After that, $\mathbf{x}_t^{\text{Tr}}$ for all t are clustered into the G group by using K-means clustering with L1 distance [8]. Consequently, the g-th group of the clustered spectrogram patch, $\mathbf{x}_g^{\text{Di}}$, is prepared. Next, unsupervised SNMF [9] is conducted on each $\mathbf{x}_g^{\text{Di}}$ to represent the g-th group with R spectral bases, $\mathbf{b}_g^{\text{x}}$. Finally, $\mathbf{b}_g^{\text{x}}$ of G groups concatenate together to form the dictionary corresponding to the target sound event as

$$\mathbf{B}_x = \left[\mathbf{b}_{x;1} \cdots \mathbf{b}_{x;g} \cdots \mathbf{b}_{x;G}\right]. \tag{2}$$

2.2. SNMF-based source separation

In the source separation step, the mixture of the target sound event and interfering noise signals is given as

$$y_t(n) = x_t(n) + d_t(n) \tag{3}$$

where $x_i(n)$ and $d_i(n)$ are the target sound event and the interfering noise at the t-th frame, respectively. By applying a short-time Fourier transform (STFT) and the STFT-to-mel conversion matrix to (3), $y_t(n)$ can be represented in the frequency domain as

$$|Y_t(k)|^2 \cong |X_t(k)|^2 + |D_t(k)|^2 \text{ for } k = 0, \cdots, K-1 \tag{4}$$

where $|Y_t(k)|^2$, $|X_t(k)|^2$, and $|D_t(k)|^2$ denote the k-th mel-spectral powers of $y_t(n)$, $x_t(n)$, and $d_t(n)$, respectively. Similar to the dictionary training stage, $Y_t(k)$ is also represented in the spectrogram patch, $\mathbf{Y}_t$, as

$$\mathbf{y}_t = \{|Y_{t-P+1}(k)|^2, \cdots, |Y_t(k)|^2\}. \tag{5}$$

In the NMF framework, $\mathbf{y}_t = \mathbf{B}_y \mathbf{A}_{y;t}$, $\mathbf{x}_t = \mathbf{B}_x \mathbf{A}_{x;t}$, and $\mathbf{d}_t = \mathbf{B}_{d;t} \mathbf{a}_{d;t}$, where $\mathbf{B}_y$, $\mathbf{B}_x$, and $\mathbf{B}_{d;t}$ are the dictionaries of $\mathbf{y}_t$, $\mathbf{x}_t$, and $\mathbf{d}_t$, respectively. Moreover, $\mathbf{A}_{y;t}$, $\mathbf{A}_{x;t}$, and $\mathbf{A}_{d;t}$ are the activation matrices corresponding to $\mathbf{B}_y$, $\mathbf{B}_x$, and $\mathbf{B}_{d;t}$ at the t-th frame, respectively. By assuming that $\mathbf{x}_t$ and $\mathbf{d}_t$ are fully separable from $\mathbf{y}_t$, $\mathbf{y}_t$ can be rewritten as [11]

$$\mathbf{y}_t = \mathbf{B}_y \mathbf{A}_{y;t} = \left[\mathbf{B}_x \mathbf{B}_{d;t}\right] \begin{bmatrix} \mathbf{A}_{x;t} \\ \mathbf{A}_{d;t} \end{bmatrix} = \mathbf{B}_x \mathbf{A}_{x;t} + \mathbf{B}_{d;t} \mathbf{A}_{d;i} \tag{6}$$

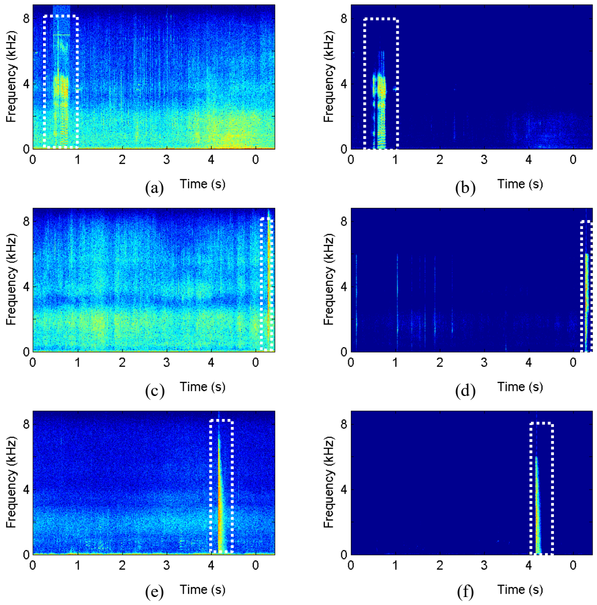
where $\mathbf{B}_y = \left[\mathbf{B}_x \mathbf{B}_{d;t}\right]$ and $\mathbf{A}_{y;t} = \left[\mathbf{A}_{x;t} \mathbf{A}_{d;t}\right]^{\text{T}}$. Note that T refers to the transpose operation. If R_x and R_d ($R_y = R_x + R_d$) are the ranks of the dictionaries for $\mathbf{x}_t$ and $\mathbf{d}_t$, respectively, then the dimensions of $\mathbf{B}_y$, $\mathbf{B}_x$, and $\mathbf{B}_{d;t}$ are $K \times R_y$, $K \times R_x$, and $K \times R_d$, respectively, while the dimensions of $\mathbf{A}_{y;t}$, $\mathbf{A}_{x;t}$, and $\mathbf{A}_{d;i}$ are $R_y \times P$, $R_x \times P$, and $R_d \times P$, respectively. Note that R_x and R_d are set to RG in this paper.

Since supervised NMF assumes that both $\mathbf{B}_x$ and $\mathbf{B}_{d;t}$ are given in advance [7], they focus on finding $\mathbf{A}_{x;t}$ and $\mathbf{A}_{d;t}$ from $\mathbf{y}_t$ for the separation of speech and noise. To achieve this goal, a multiplicative update rule with a sparsity constraint [9] is iteratively performed as

$$\begin{bmatrix} \mathbf{A}_{x;t}^{j} \\ \mathbf{A}_{d;t}^{j} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{x;t}^{j-1} \\ \mathbf{A}_{d;t}^{j-1} \end{bmatrix} \otimes \frac{\left[\mathbf{B}_x \mathbf{B}_{d;t}\right]^{\text{T}} \dfrac{\mathbf{y}_i}{\left[\mathbf{B}_x \mathbf{B}_{d;t}\right]\left[\mathbf{A}_{x;t}^{j-1} \mathbf{A}_{d;t}^{j-1}\right]^{\text{T}}}}{\left[\mathbf{B}_x \mathbf{B}_{d;t}\right]^{\text{T}} \mathbf{1} + \boldsymbol{\mu}} \tag{7}$$

where j is an iteration index and $\boldsymbol{\mu}$ is an $R_y \times 1$ matrix in which all elements are equal to a sparsity weight of the ℓ_1 constraint, which is set to 5 according to the previous work [9]. In addition, $\otimes$ and / indicate element-wise multiplication and division, respectively. Moreover, $\mathbf{1}$ in (4) is a $K \times 1$ matrix in which all elements are equal to unity. Note that all the elements of $\mathbf{a}_{y;i}^{j=0} = \left[\mathbf{A}_{x;t}^{j=0} \mathbf{A}_{d;t}^{j=0}\right]^{\text{T}}$ can be initialized as random values between 0 and 1 [9]. In NMF separation, (4) is repeated until the relative reduction of an NMF objective function is less than a pre-defined threshold. In this paper, the Kullback–Leibler (KL) divergence is employed as an NMF objective function [7, 9].

After the source separation, IRM is estimated from $\mathbf{B}_x$, $\mathbf{B}_{d;t}$, $\mathbf{A}_{x;t}$, and $\mathbf{A}_{d;t}$ by using MMSE filtering [7] for both the feature enhancement and noise estimation for ONL. To this end, the *a*



(a) (b) (c) (d) (e) (f)

Figure 2: Spectrogram comparisons for noisy mixture and estimated target event by the proposed source separation: left and right columns show spectrograms of noisy mixtures and the separated sound events for baby-cry, glass-crash, and gunshot classes, respectively.

priori SNR, $\boldsymbol{\xi}_t$, is first estimated with a decision-directed approach as

$$\boldsymbol{\xi}_t = \frac{\alpha \tilde{\mathbf{x}}_{t-1} + (1-\alpha)\mathbf{B}_x \mathbf{A}_{x;t}}{\bar{\mathbf{d}}_{t-1}} \tag{8}$$

where α is a smoothing coefficient for the decision-directed ξ_i. In addition, $\bar{\mathbf{d}}_i$ in (7) is a time-smoothed noise estimate of $\mathbf{B}_{d;t}\mathbf{A}_{d;t}$, and it is realized as

$$\bar{\mathbf{d}}_t = \gamma \bar{\mathbf{d}}_{t-1} + \beta_t (1-\gamma)\mathbf{B}_{d;t}\mathbf{A}_{d;t} \tag{9}$$

where $\bar{\mathbf{d}}_0 = \hat{\mathbf{d}}_1$, and γ controls the stationarity of $\bar{\mathbf{d}}_i$. In (9), β_i is an adaptive noise flooring factor at the *i*-th frame, which is derived from the ratio between the normalized activation powers of the separated noise and speech as

$$\beta_t = 20\log_{10}\frac{R_x \sum_{r=1}^{R_d} A_{d;t}(r)}{R_d \sum_{r=1}^{R_x} A_{x;t}(r)} \tag{10}$$

where $A_{x;t}(r)$ and $A_{d;t}(r)$ indicate an *r*-th element of $\mathbf{A}_{x;t}$ and $\mathbf{A}_{d;t}$ from (7), respectively. Next, the IRM is constructed as

$$\mathbf{m}_t = \frac{\boldsymbol{\xi}_t}{1+\boldsymbol{\xi}_t} \tag{11}$$

and an enhanced spectrogram patch of the sound event, $\tilde{\mathbf{x}}_t$, is obtained by applying (11) to $\mathbf{y}_t$; thus, $\tilde{\mathbf{x}}_t = \mathbf{m}_t \otimes \mathbf{y}_t$. Finally, $\tilde{\mathbf{x}}_t$ is flattened into a *KP* vector and is then fed into the DNN-based binary classifier as an input feature to determine whether the event has occurred at the *t*-th frame.

TABLE I

PARAMETERS SETTING FOR THE PROPOSED SOURCE SEPARATION.

Parameters	Description	Value
K	Number of mel-spectral bins	64
P	Time frames consisting a spectrogram patch	12
R	Number of bases consisting a group	6
G	Number of group consisting a dictionary	20

2.3. Online noise learning

To cope with various unseen noise on-the-fly, the proposed ONL first estimates a mel-spectral power of the reference noise, $\tilde{\mathbf{d}}_t$, by using the IRM, $\mathbf{m}_t$, as described in (11). That is, $\tilde{\mathbf{d}}_t$ is estimated only when the noise activation is dominant, such as

$$\tilde{\mathbf{d}}_t = \begin{cases} \mathbf{y}_t \otimes (1-\mathbf{m}_t), & if \quad \beta_t > 0 \\ \bar{\mathbf{d}}_t, & otherwise \end{cases} \tag{12}$$

In this work, each noise basis is tested for whether it should be updated by

$$I_t(r) = \begin{cases} 1, & if \;\; A_{d;\hat{t}}(r) > \bar{A}_t \\ 0, & otherwise \end{cases} \tag{13}$$

where $\bar{\mathrm{A}} = (\sum_{r=1}^{R_x} A_{x;t}(r))/R_x$ and $I(r) = 1$ means that the *r*-th basis should be updated to accommodate the noise that appears at the *t*-th frame. Then, $\mathbf{A}_{d;t}$ is decomposed depending on (13) into $\mathbf{A}_{d;t}^{r\in \boldsymbol{I}_u}$ and $\mathbf{A}_{d;t}^{r\in \boldsymbol{I}_f}$, where $\boldsymbol{I}_u = \{r|I(r)=1\}$ and $\boldsymbol{I}_f = \{r|I(r)=0\}$. By using $\tilde{\mathbf{d}}_t$ and $\mathbf{A}_{d;t}^{r\in \boldsymbol{I}_u}$, the learnt noise dictionary for the (*i*+1)-th frame, $\widehat{\mathbf{B}}_{d;t+1}^{j}$, is iteratively updated by using the SNMF that minimizes KL divergence [9] as

$$\widehat{\mathbf{B}}_{d;t+1}^{j} = \widehat{\mathbf{B}}_{d;t+1}^{j-1} \otimes \frac{\dfrac{\tilde{\mathbf{d}}_t}{\widehat{\mathbf{B}}_{d;t+1}^{j-1}\left(\mathbf{A}_{d;t}^{r\in \boldsymbol{I}_u}\right)^{\mathrm{T}}}\left(\mathbf{A}_{d;t}^{r\in \boldsymbol{I}_u}\right)^{\mathrm{T}}}{\mathbf{1}\left(\mathbf{A}_{d;t}^{r\in \boldsymbol{I}_u}\right)^{\mathrm{T}}} \tag{14}$$

where $\widehat{\mathbf{B}}_{d;t+1}^{j=0} = \mathbf{B}_{d;t}^{r\in \boldsymbol{I}_u}$ and *j* is an iteration index. Finally, $\mathbf{B}_{d;t+1}$ is obtained by concatenating the converged $\widehat{\mathbf{B}}_{d;t+1}^{j^*}$ and fixed noise dictionary, $\mathbf{B}_{d;t}^{r\in \boldsymbol{I}_f}$, as $\mathbf{B}_{d;t+1} = [\widehat{\mathbf{B}}_{d;t+1}^{j^*}\, \mathbf{B}_{d;t}^{r\in \boldsymbol{I}_f}]$, which will be used for the SNMF-based sound event and noise separation for the next frame.

Figure 2 demonstrates the proposed source separation in spectrogram comparisons in a linear frequency scale. As indicated in the figure, each target sound event was successfully separated from unseen noises with diverse levels and characteristics.

3. PERFORMANCE EVALUATION

To evaluate the performance of the proposed source separation method for the SED task, we conducted rare sound event detection experiments that were provided by the IEEE DCASE 2017 Challenge: Task 2 [10]. The training dataset was composed of the background recordings from the TUT Acoustic Scenes 2016 development dataset [12] and the three classes of rare sound events: baby-cry, glass-break, and gunshot. Both the background recordings and event classes are mixed by the mixture generation procedure with the default recipe provided by DCASE 2017 Task 2, except the total number of development training sets was

TABLE II
PERFORMANCE RESULTS FOR DCASE 2017 TASK 2 FOR THE DEVELOPMENT SET, EVENT-BASED METRICS

Event Classes	DNN Classifier		Proposed Source Sep. + DNN Classifier	
	F1	ER	F1	ER
baby-cry	72.93 %	0.68	74.87 %	0.60
glass-break	87.08 %	0.26	92.34 %	0.15
gunshot	57.14 %	0.70	63.26 %	0.64
Average	72.38 %	0.54	76.94 %	0.46

increased from 500 to 1500 for the DNN training [10]. Note that the proposed source separation was employed for both the training and the test mixture data for the DNN training.

The parameters regarding the proposed method were set as shown in Table I. In addition, the frame length and shift for the STFT were set to 40 and 20 ms, respectively. The STFT-to-mel conversion matrix was designed so that 64 mel-spectral bands ranged between 80 Hz and 15 kHz. Note that all the input features were normalized to compensate for the power mismatch between data. The dictionary for the proposed method was trained according to Section 2.1 with a training set for each event. Note that the initial noise dictionary, $\mathbf{B}_{d;t=0}$, was also trained in advance with the training dataset from the background recordings.

The DNN-based event classifier was developed as follows. The input feature vector consisted of five consecutive frames, which were each represented as 64-dimensional log mel-spectral power, resulting in 320 visible units being used as an input layer. The DNN had three hidden layers with hidden rectified linear units of 256, 128, and 64, respectively. The output layer had one unit with the sigmoid activation function, corresponding to the occurrence probability of the target sound event. The parameters of the network were initialized by random values generated from uniform distribution. The fine-tuning of the network was performed using binary cross entropy as the loss function through error back propagation supervised by the correct labeling of frames. The mini-batch size for the stochastic gradient descent algorithm was set as 1024. The training was stopped after 200 epochs. The dropout percentage of 20% was applied for regularization. In the post-processing stage, frame binarization with a threshold of 0.5 and median filtering with a window length of 0.54 s were applied to smooth out the detection results. Note that the DNNs corresponding to three different event classes were trained separately.

As evaluation metrics, the F-score and error rate (ER) were used on the event-based metric [13]. The F-score and ER were measured by the *sed_eval* toolbox provided by DCASE 2017. Table II shows the evaluation results of the DNN-based event classifiers employing the proposed source separation. As indicated in the table, the proposed source separation resulted in performance improvement over the DNN classifier trained with noisy mixtures in terms of both the F-score and ER for different even classes. On average, the proposed source separation relatively increased the F-score by 6% and decreased the ER by 14.81%.

4. CONCLUSIONS

In this paper, the source separation method using SNMF with ONL is proposed to improve the detection accuracy of rare sound event detection in unseen noise conditions. The proposed method separates the mel-spectral power of the target sound event from the noisy mixture using the supervised NMF and updates the noise dictionary on-the-fly by using the separation results of subsequent frames. The separated mel-spectral power was fed as a feature vector of the DNN-based binary classifier. It was shown from the experiment provided by DCASE 2017: Task 2 that the proposed method achieved a higher average F-score and lower average ER than a DNN-based binary classifier without employing the source separation method.

5. ACKNOWLEDGMENT

Author would like to thanks to Toni Heittola, Aleksandr Diment, and Annamaria Mesaros for providing the baseline and dataset of the DCASE 2017 Challenge: Task 2.

6. REFERENCES

[1] M. Cristani, M. Bicego, and V. Murino, “Audio-visual event recognition in surveillance video sequences,” *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257-267, Feb. 2007.

[2] J. F. Gemmeke, L. Vuegen, P. Karsmakers, and B. Vanrumste, “An exemplar-based NMF approach to audio event detection,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, doi:10.1109/WASPAA.2013.6701847.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2007, pp. 21-26.

[4] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, “Sound event detection in multisource environments using source separation,” in *Proc. Workshop on Machine Listening in Multisource Environments*, 2011, pp. 36-40.

[5] K. M. Jeon, D. Y. Lee, H. K. Kim, and M. J. Lee, “Acoustic surveillance of hazardous situations using nonnegative matrix factorization and hidden Markov model,” in *Proc. Audio Engineering Society (AES) 137th Convention*, 2014, Preprint 9203.

[6] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, “Acoustic event detection method using semi-supervised nonnegative matrix factorization with a mixture of local dictionaries,” in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[7] K. M. Jeon and H. K. Kim, “Local sparsity based online dictionary learning for environment adaptive speech enhancement with nonnegative matrix factorization,” in *Proc. Interspeech*, 2016, pp. 2861-2865.

[8] D. Arthur and S. Vassilvitskii, “K-means++: the advantages of careful seeding,” in *Proc. Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027-1035.

[9] J. Le Roux, F. J. Weninger, and J. R. Hershey, *Sparse NMF–half-baked or Well Done?*, Mitsubishi Electric Research Labs (MERL), Cambridge, MA, Tech. Rep. TR-2015-23, 2015.

[10] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. “DCASE 2017 challenge setup: tasks, datasets and baseline system,” in *Proc. Work-*

shop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2017, submitted.

[11] K. M. Jeon, H. K. Kim, S. J. Lee, and Y. K. Lee, "Nonnegative matrix factorization based adaptive noise sensing over wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 10, no. 4, doi:10.1155/2014/640915, Jan. 2014.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.