ACOUSTIC EVENT DETECTION BASED ON NON-NEGATIVE MATRIX FACTORIZATION WITH MIXTURES OF LOCAL DICTIONARIES AND ACTIVATION AGGREGATION

Tatsuya Komatsu, Yuzo Senda, and Reishi Kondo

Information and Media Processing Laboratories
NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki 211-8666, Japan

ABSTRACT

This paper proposes a new non-negative matrix factorization (NMF) based acoustic event detection (AED) method with mixtures of local dictionaries (MLD) and activation aggregation. One of the key problems of conventional NMF-based methods is instability of activations due to redundancy of a region spanned by the bases of dictionaries. Sounds inside the redundant region are often decomposed into undesired combinations of bases and activations that cause failure of detection. The proposed method employs MLD for allocating sub-groups of basis dictionaries to acoustic elements to minimize redundancy in the region and obtain controlled activations. In order to make activations more stable, the proposed method also introduces activation aggregation which combines basis-wise activations into acoustic-element-wise activations. Much more stable activations by the proposed method lead to significant improvement in F-measure by up to 60% compared to an ordinary convolutive-NMF-based method. The proposed method also outperforms a latest alternative which is not based on NMF.

Index Terms— Acoustic Event Detection, NMF, group sparsity, convex cone, activation aggregation

1. INTRODUCTION

To make cities safer, acoustic event detection (AED) as part of a monitoring system is expected to find hazardous sounds related to crimes, accidents and incidents in public spaces. In the spaces, environmental sounds coexist with the target sounds. This results in failure of detection. Non-negative matrix factorization (NMF) has been studied as a blind source separation method for this kind of situation[1, 2]. A number of AED methods based on NMF have been proposed[3, 4, 5, 6]. However, there is still room for improvement as indicated by a comparison[7]. One of the key problems of NMF-based methods is instability of activations due to overlaps and unnecessary coverages of a region spanned by a basis dictionary. Unstable activations can adversely affect the performance of the subsequent classifier which uses activations as feature vector.

An acoustic event consists of a series of physical phenomena, e.g., a glass break event starts from an impact to the glass followed by resonances of broken pieces. In this paper, an acoustic element refers a short sound produced by a physical phenomenon. Conventional NMF-based methods[7, 8] model the acoustic elements by a basis dictionary to decompose acoustic events into a combination of bases and corresponding weights/activations. The subsequent classifier uses the activations as a feature vector representing the content rate of the corresponding acoustic element. However, those methods often fail to capture the acoustic elements and excite unstable activations due to their redundancy of regions spanned by dictionaries. From the perspective of geometrical NMF interpretations[9, 10], the redundant region can be interpreted as overlaps or unnecessary coverages of convex cones formed by dictionaries. To reduce such redundancy and capture the acoustic elements correctly, Mixtures of Local Dictionaries (MLD) [11] is a promising NMF method when allocating a sub-group of basis dictionary to each acoustic element. Each basis group forms a small convex cone controlled to be as small as possible to prevent overlap between cones and unnecessary coverages in a dictionary. Introducing sparseness among the cones helps NMF decompose the mixture into spaces of the cones.

This paper proposes a new AED method based on convolutive NMF with MLD. As the cones are linked to the acoustic elements, AED needs to know assumedly only the activation for each cone rather than for each basis. Based on this assumption, activation aggregation is newly introduced to make activations more stable. Since different events may have the same acoustic element, an event should be classified according to the combination of its acoustic elements. The proposed method is evaluated with a variety of environments and noise levels to confirm its robustness.

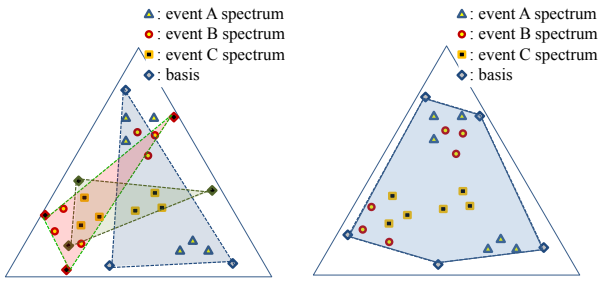
2. PROBLEM IN CONVENTIONAL METHODS

J. F. Gemmeke et al. [8] employs a compositional model to make a dictionary matrix $\mathbf{W}$ from event specific basis matrices $\mathbf{W}_{(i)}$ where $i \in \{1, ..., I\}$ represents an event index. Each $\mathbf{W}_{(i)}$ is extracted by performing NMF to an event specific spectrogram $\mathbf{V}_{(i)}$ individually. In the classification phase, an unclassified spectrogram $\mathbf{V}_{(*)}$ is decomposed by NMF with the dictionary $\mathbf{W}$. The resulting activation matrix consists of event specific activation matrix $\mathbf{H}_{(i)}$ corresponding to $\mathbf{W}_{(i)}$ as follows:

$$\mathbf{V}_{(*)} \sim \left[\mathbf{W}_{(1)}, ..., \mathbf{W}_{(I)}\right] \begin{bmatrix} \mathbf{H}_{(1)} \\ \vdots \\ \mathbf{H}_{(I)} \end{bmatrix}. \quad (1)$$

C. V. Cotton et al. [7] applies NMF to the entire training data spectrogram $\mathbf{V}$ made by concatenating event specific spectrograms $\mathbf{V}_{(i)}$. In this way of dictionary generation, the extracted basis matrix $\mathbf{W}$ represents a universal basis matrix for describing any acoustic events, thus each basis does not refer to any particular event but an acoustic element. Hence, NMF is performed to each event specific spectrogram $\mathbf{V}_{(i)}$ using the universal dictionary $\mathbf{W}$ to obtain an event specific $\mathbf{H}_{(i)}$.

$$\mathbf{V}_{(i)} \sim \mathbf{W}\mathbf{H}_{(i)}. \quad (2)$$



(a) J. F. Gemmeke et al.[8] (b) C. V. Cotton et al.[7]

Fig. 1. Dictionaries of conventional works.

$\mathbf{H}_{(i)}$ represents a content ratio of acoustic elements in $\mathbf{V}_{(i)}$. HMM is employed to detect events from the resulting activation matrix of an unclassified spectrogram.

These methods seem to tie activations and events successfully; however, they overlook the nature of NMF. A combination of basis vectors forms a convex cone that can reconstruct any data points inside the cone. Once data points are projected onto the simplex, the cone can be discussed as a convex hull (see 6.2). Fig. 1 illustrates geometrical interpretations of decomposition in the conventional methods. The triangle in the figure represents the simplex of the orthant where data points exist. Basis vectors are on the simplex by definition and data points are projected onto the simplex. Gemmake's dictionary shown in Fig. 1(a) has convex cones, each of which spans widely to enclose corresponding event's data points. There are overlaps among the convex hulls since the events tend to have the same acoustic elements. Data points in the overlaps may excite unstable activations, which result in failure of classification. In Fig. 1(b), Cotton's dictionary is expressed as a single big convex hull since it is a universal basis matrix. There could be a number of combinations to represent a data point inside the cone. In other words, data points produce unexpected patterns of content ratio. The patterns cause false positive and false negative in HMM detection.

3. THE PROPOSED METHOD

A combination of convolutive NMF, MLD, activation aggregation and support vector machine (SVM) is proposed here as an AED method. The proposed method consists of two main parts, i.e. dictionary generation and event classification. The dictionary generation relys on MLD to eliminate overlaps among convex cones and unnecessary coverages in a dictionary. The event classification carrys out three stages, NMF with the dictionary, activation aggregation and SVM classification. Fig. 3 shows the entire block diagram of the proposed method.

3.1. Dictionary Generation

The basis matrices for G small convex cones are concatenated to form an MLD dictionary $\mathbf{W} = [\mathbf{W}^{(1)}, ..., \mathbf{W}^{(G)}]$. A basis matrix $\mathbf{W}^{(g)} \in \mathcal{R}_{+}^{F \times K_g}$ consists of K_g basis vectors where $g \in \{1, ..., G\}$ is an index to each acoustic element. To determine target acoustic elements, a universal basis matrix $\mathbf{W}_0$ is first extracted from the entire training data spectrogram $\mathbf{V}$ with an ordinary convolutive NMF. K-means clustering is then applied to the basis vectors in the matrix to select G centroids $\boldsymbol{\mu}^{(g)}$ as the targets. Convolutive NMF is applied again to $\mathbf{V}$ with the centroids $\boldsymbol{\mu}^{(g)}$ and the following cost function

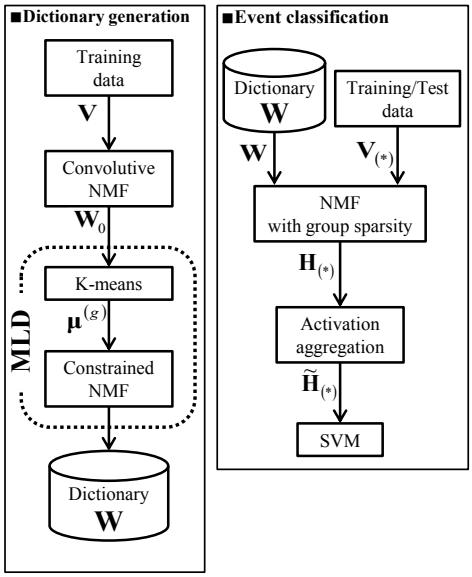


Fig. 2. Block diagram of the proposed acoustic event detection.

using the generalized Kullback-Leibler (KL) divergence $\mathcal{D}_{\mathcal{KL}}(\cdot|\cdot)$.

$$\mathcal{D}(\mathbf{V}|\mathbf{\Lambda}) = \mathcal{D}_{\mathcal{KL}}(\mathbf{V}|\mathbf{\Lambda}) + \eta \sum_{g} \mathcal{D}_{\mathcal{KL}}(\boldsymbol{\mu}^{(g)}|\mathbf{W}^{(g)}) + \lambda \sum_{t} \Omega(\mathbf{h}_t) \quad (3)$$

where $\mathbf{\Lambda} = \mathbf{WH}$, $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_T]$ and $\mathbf{h}_t = [\mathbf{h}_t^{(1)}, ..., \mathbf{h}_t^{(G)}]^\top$ and $(\cdot)^\top$ denotes the transpose of the matrix.

The second term leads to a small convex cone enclosed by $\mathbf{W}^{(g)}$ for the gth acoustic element characterized by the centroid $\boldsymbol{\mu}^{(g)}$. The size of a cone is controlled by η. The third term represents group activation sparsity at time t controlled by λ. In this article, we borrow $\Omega(\mathbf{h}_t) = \sum_g \log(\epsilon + \|\mathbf{h}_t^{(g)}\|_1)$, from a prior art[11, 12].

As MLD approach is applied to convolutive NMF, the update rule is modified to:

$$\mathbf{W}_\theta^{(g)} \leftarrow \mathbf{W}_\theta^{(g)} \odot \left\{ \left(\frac{\mathbf{V}}{\mathbf{\Lambda}} \right) \overset{\theta\rightarrow}{\mathbf{H}}^\top + \eta \frac{\boldsymbol{\mu}_\theta^{(g)}}{\mathbf{W}_\theta^{(g)}} \right\} \Big/ \left\{ \mathbf{1} \left(\overset{\theta\rightarrow}{\mathbf{H}}^\top + \eta \right) \right\}, \quad (4)$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left\{ \sum_{\theta=0}^{\Theta-1} \mathbf{W}_\theta^\top \left(\frac{\overset{\theta\leftarrow}{\mathbf{V}}}{\mathbf{\Lambda}} \right) \right\} \Big/ \left\{ \sum_{\theta=0}^{\Theta-1} \mathbf{W}_\theta^\top \mathbf{1} \right\}, \quad (5)$$

$$\mathbf{h}_t^{(g)} \leftarrow \mathbf{h}_t^{(g)} \frac{1}{1 + \lambda \Big/ \left\{ \epsilon + ||\mathbf{h}_t^{(g)}||_1 \right\}} \quad (6)$$

where $\odot$ and $\Big/$ represent the element wise multiplication and division and $\mathbf{W}_\theta$ indicates a basis matrix over a plurality of frames with a frame length $\theta \in \{0, ..., \Theta - 1\}$, and $\overset{\theta\rightarrow}{(\cdot)}$ is a column shift operator described in [13].

Fig. 3.1 explains the relationship among data points, basis vectors and convex cones generated by this process. Each convex cone spans minimally enough to enclose data points of an acoustic element.

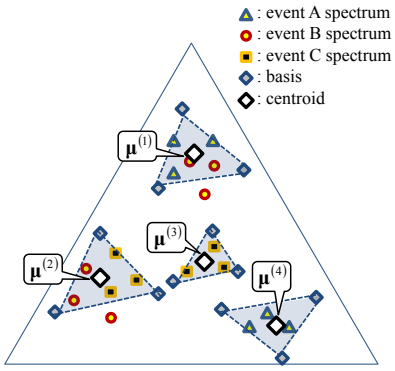


Fig. 3. Dictionary of the proposed method.

3.2. Event Classification

A target spectrogram is decomposed into an activation matrix by NMF with the MLD dictionary generated above.

$$\mathbf{V}_{(i)} \sim \mathbf{W}\mathbf{H}_{(i)}. \tag{7}$$

$\mathbf{H}_{(i)}$ is updated using Eq. (5) and group sparsity Eq. (6) which helps $\mathbf{H}_{(i)}$ to be sparse by turning off activations of the irrelevant acoustic elements.

However, there is still instability in $\mathbf{h}_t^{(g)}$, an activation of g th group. Therefore, activation aggregation $\tilde{h}_t^{(g)} = ||\mathbf{h}_t^{(g)}||_1$ is introduced here to combine the basis-wise activations into $\tilde{\mathbf{h}}_t = [\tilde{h}_t^{(1)}, ..., \tilde{h}_t^{(g)}]^\top$. As a convex cone spanned by $\mathbf{W}^{(g)}$ is linked to g th acoustic element, $\tilde{h}_t^{(g)}$ can be treated as a acoustic-element-wise activations. This activation aggregation makes activations more stable as a feature vector and helps the subsequent classifier to learn the relationship between activations and events.

In the training phase, an event specific aggregated activation $\tilde{\mathbf{H}}_{(i)} = [\tilde{\mathbf{h}}_{(i),1}, ..., \tilde{\mathbf{h}}_{(i),T}]$ is extracted from the corresponding spectrogram $\mathbf{V}_{(i)}$. Once $\tilde{\mathbf{H}}_{(i)}$ are obtained, they are used as feature vectors to train the SVM. In this article, a simple linear SVM is used.

4. EXPERIMENTS

Experiments were performed on synthetic data to evaluate AED performance of the proposed method. For comparison, we also experimented two conventional methods. One is the convolutive NMF dictionary method[7] with sparseness constraint [14]. The difference between the proposed method and convolutive NMF are MLD procedure and activation aggregation. The other conventional method is a latest alternative which is non-NMF-based method proposed by X. Lu et al.[15].

4.1. Latest alternative not based on NMF

To the best of our knowledge, Lu's method is a good alternative among latest methods not limited to NMF based. The method, inspired by a work in image processing, employs a bag of spectral patch exemplars to capture the temporal-frequency structures of acoustic events. The method whitens the patches and then finds representative patches by applying k-means clustering. The feature vector of a patch is extracted by lining up the similarity measure for each representative patch. With this feature, an SVM classifier is built for AED. This method is not based on NMF, but similar ideas can be found in it.

Table 1. Test environments.

Environment	Background sound sources
station square	speech, music,
station concource	speech, announcement, footsteps, train, chime
airport lobby	speech, announcement, footsteps, cart, suitcase
bus terminal	bus buzzer, engine, footsteps, announcement
suburb	insects, river, car engine (truck)

4.2. Experimental condition

Assuming an application to city monitoring, tasks are set to detection of three major events, scream, glass break, and gunshot, in five environments. Sound data sets used are Series 6000 General Sound Effects Library[16] and ATR environmental dataset [17]. Table 1 shows detail of the environments. Test data were synthesized by mixing a clean event sound signal into an environmental sound signal with signal-to-noise ratio (SNR) controled at 10, 15, and 20 dB. All the signals are 16 kHz sampling and applied FFT with 512 pt frame and 256 pt shift. Acoustic event detection performance is measured by frame-wise recall, precision and F-measure[18] of 5-fold cross validation.

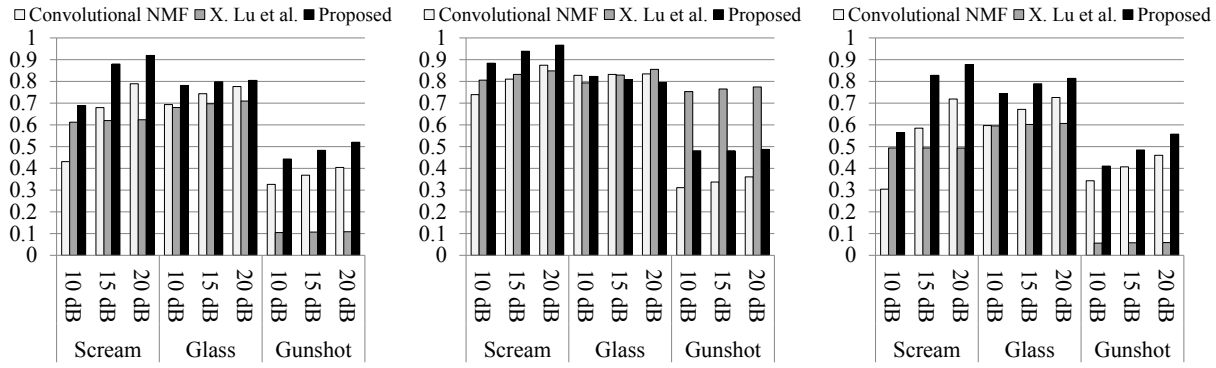
The dictionaries were learned with a size of group $G = 30$ which holds $K_g = 5$ bases (total: 150 bases) of frame size $\Theta = 5$ for the proposed method. For the convolutive NMF method, 150 bases, which is equivalent to total size of the proposed method, and $\Theta = 5$. Lu's method uses codebook size 128 and patch size 10. Other parameters of all the methods were set to optimal values derived by preliminary experiments.

4.3. Experimental results and discussion

Fig. 4 shows the overall performance in five environments. The proposed method outperformed both the conventional methods among all SNRs. Especially at 'scream' in 10 dB, the proposed method showed 60% improvement in F-measure compare to the convolutive NMF.

Some remarkable differences to the proposed method tell flaws of the conventional methods. The precision differences for scream indicate that convolutive NMF detected a lot of false positives. This seems because its big convex cone includes unnecessary regions where no target exists but environmental sounds got into and excited similar patterns of activations.

As expected, Lu's method keeps the point as it worked well at a low SNR of 10dB. However, it could not gain much improvement in recall even at 20dB. Simplicity of the method, a good point of this method, may result in limitation of capability. The results for gunshot reveal the weak point of Lu's method. Since its codebook can hardly express impulse sounds due to its normalization process, sounds not resembling any word in it may be detected as a gunshot event.



(a) F-measure (b) Recall (c) Precision

Fig. 4. The overall performance in five environments.

5. CONCLUSIONS

This paper has proposed a new AED method based on convolutive NMF with MLD dictionary generation and activation aggregation. The proposed method employed MLD for allocating small convex cones to acoustic elements to minimize overlaps and coverages in regions spanned by the dictionary. Since MLD's cost function led to group sparsity in basis as well as in activation, the function was also applied to the event classification. To make activations more stable, the proposed method also introduced activation aggregation which combines basis-wise activations into acoustic-element-wise activations. Compared to ordinary convolutive NMF, much more stable activations by the proposed method led to significant improvement in F-measure by up to 60%. The proposed method also outperformed a latest alternative which is not based on NMF.

6. APPENDIX

6.1. Non-Negative Matrix Factorization

To understand the nature of NMF, a basic algorithm and a geometrical interpretation are revisited. In the AED context, NMF is an algorithm to find a pair of matrices such that their product $\mathbf{\Lambda}$ approximates a source spectrogram $\mathbf{V} \in \mathbb{R}_+^{F \times T}$,

$$\mathbf{V} \sim \mathbf{\Lambda} = \mathbf{WH}, \tag{8}$$

where $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ is a basis matrix and $\mathbf{H} \in \mathbb{R}_+^{K \times T}$ is a activation matrix[19]. $\mathbf{W}$ and $\mathbf{H}$ are estimated to minimize an cost function $\mathcal{D}\left(\mathbf{V}|\mathbf{\Lambda}\right)$. The popular choice of cost function is the generalized KL divergence,

$$\mathcal{D}_{\mathcal{KL}}\left(x|y\right) = x \log \frac{x}{y} - x + y. \tag{9}$$

To minimize $\mathcal{D}_{\mathcal{KL}}\left(\mathbf{V}|\mathbf{\Lambda}\right)$, $\mathbf{W}$ and $\mathbf{H}$ are updated in a multiplicative way by turn as follows:

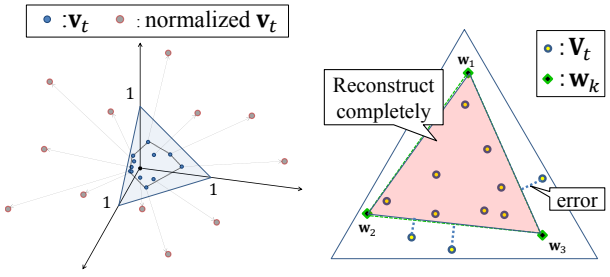
$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\mathbf{\Lambda}}\mathbf{H}^\top}{\mathbf{1}\mathbf{H}^\top}, \quad \mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top \frac{\mathbf{V}}{\mathbf{\Lambda}}}{\mathbf{W}^\top \mathbf{1}}. \tag{10}$$

where $\mathbf{1}$ is a $F \times T$ matrix with all its elements are 1.

Practically, most of NMF-based AED methods use convolutive NMF to obtain spectro-temporal information of acoustic events. In that case, a basis matrix over a plurality of frames is described as $\mathbf{W}_\theta$ with the frame index $\theta \in \{0, ..., \Theta - 1\}$ and $\mathbf{V}$ is approximated as follows:

$$\mathbf{V} \sim \mathbf{\Lambda} = \sum_{\theta=0}^{\Theta-1} \mathbf{W}_\theta \overset{\theta\rightarrow}{\mathbf{H}}, \tag{11}$$

where $\overset{\theta\rightarrow}{(\cdot)}$ is a column shift operator described in [13]. If $\Theta = 1$, convulsive NMF is equivalent to general NMF. In the other part of this article, we omit subscript θ of $\mathbf{W}_\theta$ for simplicity.



(a) Datas and pullback onto the simplex (b) Data points on the simplex

Fig. 5. Relationship between data points and basis vectors.

6.2. Geometrical Interpretation of NMF

A geometrical interpretation helps us to understand the relationship among $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$[9, 10]. A row vector $\mathbf{v}_t$ at a time frame index $t \in \{1, ..., T\}$ of $\mathbf{V}$ is described by a conical combination of $\mathbf{w}_k$ corresponding to a basis index $k \in \{1, ..., K\}$ and $h_{k,t}$ which is an activation of $\mathbf{w}_k$ at the time frame index t,

$$\mathbf{v}_t = h_{1t}\mathbf{w}_1 + h_{2t}\mathbf{w}_2 + \cdots + h_{Kt}\mathbf{w}_K. \tag{12}$$

Here, if they are normalized, Eq. 12 represents a convex combination. Fig. 5 illustrates the relationship between data points and basis vectors on the simplex. $\mathbf{v}_t$ are projected onto the simplex by normalization and $\mathbf{w}_k$ constitute a convex hull that wraps most of $\mathbf{v}_t$. Due to the non-negativity of $\mathbf{W}$ and $\mathbf{H}$, the data points in the area enclosed by $\mathbf{W}$ can be completely reconstructed by a convex combination of $\mathbf{W}$ and $\mathbf{H}$, while data points outside the area cannot.

7. REFERENCES

[1] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2013, pp. 141–145.

[2] T. Higuchi and H. Kameoka, "Unified approach for underdetermined BSS, VAD, dereverberation and doa estimation with multichannel factorial HMM," in *Signal and Information Processing (GlobalSIP), IEEE Global Conference on*, 2014, pp. 562–566.

[3] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Proc CHiME*, pp. 36–40, 2011.

[4] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, pp. 341–371. Springer, 2013.

[5] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani, and K. Kashino, "Bayesian semi-supervised audio event transcription based on markov indian buffet process," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. 2013, pp. 3163–3167.

[6] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2013.

[7] C. V. Cotton and D. P. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2011, pp. 69–72.

[8] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste and H. V. Hamme, "An exemplar-based NMF approach to audio event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2013, pp. 1–4.

[9] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Advances in neural information processing systems*, 2003.

[10] C. Bauckhage, "A purely geometric approach to non-negative matrix factorization," in *16th LWA Workshops: KDML, IR and FGWM*, 2014.

[11] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *Signal Processing Letters, IEEE*, vol. 22, no. 3, pp. 293–297, 2015.

[12] A. Lefevre, F. Bach, and C. Févotte, "Itakura-saito nonnegative matrix factorization with group sparsity," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2011, pp. 21–24.

[13] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, pp. 494–499. Springer, 2004.

[14] P. D. O'grady and B. A. Pearlmutter, "Convolutive nonnegative matrix factorisation with a sparseness constraint," in *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*. IEEE, 2006, pp. 427–432.

[15] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Sparse representation based on a bag of spectral exemplars for acoustic event detection," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. 2014, pp. 6255–6259.

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

[17] ATR, "Environment sound database," http://www.atr-p.com/products/esd.html.

[18] Temko, A., Nadeu, C., and Biel, J. I., "Acoustic Event Detection: SVM-Based System and Evaluation Setup in CLEAR ' 07," in *Multi- model technologies for perception of humans*, pp. 354-363, 2008.

[19] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.