ACOUSTIC EVENT DETECTION METHOD USING SEMI-SUPERVISED NON-NEGATIVE MATRIX FACTORIZATION WITH A MIXTURE OF LOCAL DICTIONARIES

Tatsuya Komatsu, Takahiro Toizumi, Reishi Kondo, and Yuzo Senda

Data Science Research Laboratories, NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki 211-8666, Japan

ABSTRACT

This paper proposes an acoustic event detection (AED) method using semi-supervised non-negative matrix factorization (NMF) with a mixture of local dictionaries (MLD). The proposed method based on semi-supervised NMF newly introduces a noise dictionary and a noise activation matrix both dedicated to unknown acoustic atoms which are not included in MLD. Because unknown acoustic atoms are better modeled by the new noise dictionary learned upon classification and the new activation matrix, the proposed method provides a higher classification performance for event classes modeled by MLD when a signal to be classified is contaminated by unknown acoustic atoms. Evaluation results using DCASE2016 task 2 Dataset show that F-measure by the proposed method with semi-supervised NMF is improved by as much as 11.1% compared to that by the conventional method with supervised NMF.

Index Terms— Acoustic event detection, Non-negative matrix factorization, Semi-supervised NMF, Mixture of local dictionaries

1. INTRODUCTION

To identify a physical event or a sound source by which an observed acoustic signal has been produced, acoustic event detection (AED) is studied in various research fields such as smart home systems [1, 2], environmental and ecological surveillance [3, 4], and audio and video indexing [5, 6, 7]. Particularly, to make cities safer, AED as part of a monitoring system is expected to find hazardous sounds related to crimes, accidents, and incidents in public spaces [8, 9]. Environmental sound coexisting with a target acoustic signal causes wrong feature extraction and results in failure of detection. AED methods based on non-negative matrix factorization (NMF) have been proposed as promissing solutions [10, 11, 12, 13]. On AED, NMF models an acoustic event as a combination of acoustic atoms which constitutes spectra of acoustic events. NMF-based methods learn a dictionary of acoustic atoms by decomposing training signals into their spectral bases. A signal to be classified is decomposed into bases of the dictionary and the corresponding activation matrix by supervised NMF. The extracted activation matrix represents a mixture ratio of acoustic atoms in the signal and is used as a feature vector.

One of the most important points for an NMF-based AED method is how to learn a dictionary of acoustic atoms. Gemmeke et al. [14] made a dictionary by concatenating event specific basis matrices which were extracted by performing NMF on each acoustic event individually. However, when different acoustic events share the same acoustic atoms, the dictionary becomes redundant. This redundancy prevents proper extraction of an activation matrix. Komatsu et al. [15] used a mixture of local dictionaries (MLD) [16] constituting sub-groups of bases which directly models acoustic

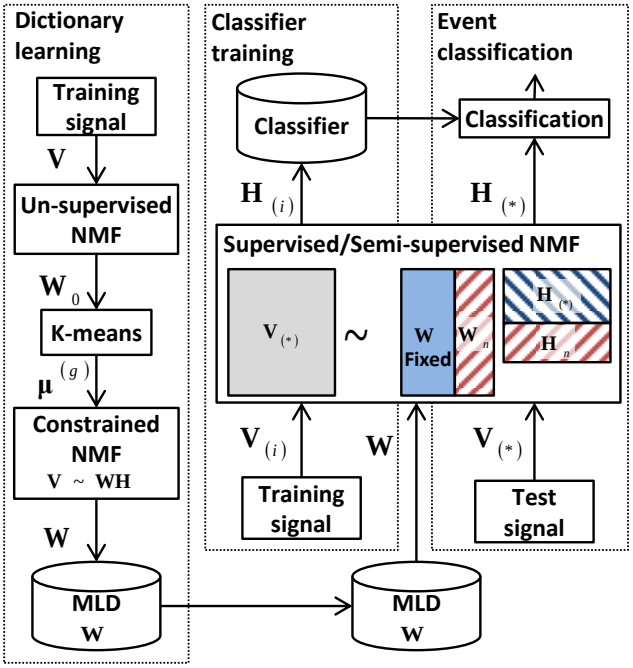

Figure 1: Block diagram of the proposed acoustic event detection.The supervised NMF is a special case of semi supervised NMF without a noise dictionary $\mathbf{W}_n$ and a noise activation matrix $\mathbf{H}_n$.

atoms. MLD is learned with constraind NMF using a prior knowledge of acoustic atoms, which is obtained from clustered spectra of training signals. Modeling acoustic atoms directly by sub-groups of basis, MLD has less redundancy and performs more accurate feature extraction. However, the conventional method performs supervised NMF [17, 18] using their fixed dictionaries upon classification. When a signal to be classified has unknown spectra (e.g. environmental sound) which are not included in training signals, the unknown spectra are expressed by acoustic atoms in the training signals. The extracted activation matrix is contaminated by unknown spectra and leads to failure of detection.

This paper proposes an AED method using semi-supervised NMF with MLD. The proposed method based on semi-supervised NMF newly introduces a noise dictionary and a noise activation matrix both dedicated to unknown acoustic atoms which are not included in training data. Because unknown acoustic atoms are better modeled by the new noise dictionary learned upon classification and the new activation matrix, the proposed method provides a higher

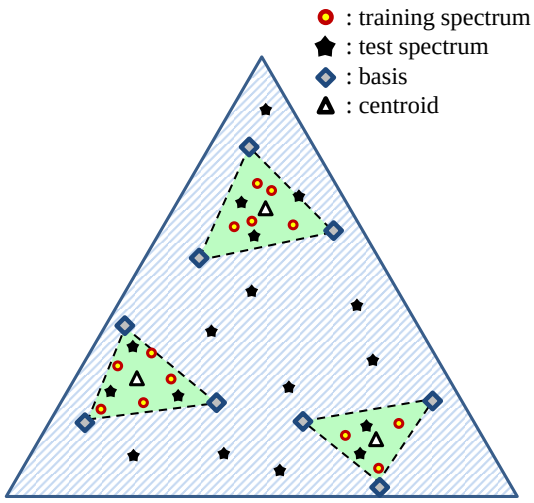

Figure 2: Relationship among training/test spectrum, MLD, and the noise dictionary

classification capability for event classes modeled by MLD when a signal to be classified are contaminated by unknown acoustic atoms.

2. PROPOSED METHOD

Figure 1 shows a block diagram of the proposed method. It consists of three parts, dictionary learning, classifier training, and event classification. Acoustic signals are used after being transformed to spectrograms.

In dictionary learning, the training spectrogram $\mathbf{V}$ is decomposed into an initial basis matrix $\mathbf{W}_0$ by basic un-supervised NMF [19]. Next, K-means clustering is applied to $\mathbf{W}_0$, and G centroids $\boldsymbol{\mu}^{(g)}$ are obtained where $g \in \{1, ..., G\}$ denotes an index of centroid. MLD $\mathbf{W}$ is learned by constrained NMF using $\boldsymbol{\mu}^{(g)}$ as prior knowledge.

In classifier training, an event-specific activation matrix $\mathbf{H}_{(i)}$ is extracted from the corresponding spectrogram $\mathbf{V}_{(i)}$ with supervised NMF using MLD $\mathbf{W}$ where i denotes an index of each acoustic event class. Column vectors of $\mathbf{H}_{(i)}$ at each time frame are used as feature vectors for training the classifier.

In event classification, unlike classifier training, semi-supervised NMF is applied to a test spectrogram $\mathbf{V}_{(*)}$ with MLD $\mathbf{W}$ and a noise dictionary $\mathbf{W}_n$ which is learned from $\mathbf{V}_{(*)}$. $\mathbf{W}_n$ and $[\mathbf{H}_{(*)}, \mathbf{H}_n]$ which are activation matrices of MLD and the noise dictionary are alternately updated. Unknown spectra included in $\mathbf{V}_{(*)}$ are expressed by $\mathbf{W}_n$ and $\mathbf{H}_n$, so that $\mathbf{H}_{(*)}$ is extracted properly. The classifier uses only $\mathbf{H}_{(*)}$ as a feature vector for classification of acoustic event classes.

2.1. Dictionary learning

MLD consists of G sub-groups of bases which model acoustic atoms $\mathbf{W} = [\mathbf{W}^{(1)}, ..., \mathbf{W}^{(G)}]$. A basis matrix $\mathbf{W}^{(g)} \in \mathcal{R}_+^{F \times K_g}$ consists of K_g basis vectors where $\mathcal{R}_+^{F \times K_g}$, F, and g denote a set of non-negative $F \times K_g$ matrices, the number of frequency bins, and an index of each acoustic atom.

To determine acoustic atoms, an initial basis matrix $\mathbf{W}_0$ is first extracted from the entire training data spectrogram $\mathbf{V} \in \mathcal{R}_+^{F \times T}$ with the basic un-supervised NMF where T denotes its number of time frames. K-means clustering is then applied to bases in $\mathbf{W}_0$ to select G centroids $\boldsymbol{\mu}^{(g)}$ which represent centroids of acoustic atoms. NMF is again applied to $\mathbf{V}$ with the centroids $\boldsymbol{\mu}^{(g)}$ and the following cost function $\mathcal{D}(\mathbf{V}|\boldsymbol{\Lambda})$:

$$\mathcal{D}(\mathbf{V}|\boldsymbol{\Lambda}) = \mathcal{D}_{\mathcal{KL}}(\mathbf{V}|\boldsymbol{\Lambda}) + \eta \sum_g \mathcal{D}_{\mathcal{KL}}(\boldsymbol{\mu}^{(g)}|\mathbf{W}^{(g)}) + \lambda \sum_t \Omega(\mathbf{h}_t), \quad (1)$$

where $\boldsymbol{\Lambda} = \mathbf{WH}$ is approximation of $\mathbf{V}$ and $\mathbf{H}$ is an activation matrix of MLD $\mathbf{W}$. A column vector $\mathbf{h}_t$ of $\mathbf{H}$ at time frame t consists of activations $\mathbf{h}_t^{(g)}$ for $\mathbf{W}^{(g)}$ (g = 1,...,G),

$$\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_t, ..., \mathbf{h}_T], \quad (2)$$

$$\mathbf{h}_t^\top = [\mathbf{h}_t^{(1)\top}, ..., \mathbf{h}_t^{(g)\top}, ..., \mathbf{h}_t^{(G)\top}], \quad (3)$$

where $[\cdot]^\top$ denotes a matrix transpose.

Cost function in (2) consists of three terms; a generalized Kullback-Leibler(KL) divergence $\mathcal{D}_{\mathcal{KL}}(\mathbf{V}|\boldsymbol{\Lambda})$ between $\mathbf{V}$ and $\boldsymbol{\Lambda}$, a constraint $\sum_g \mathcal{D}_{\mathcal{KL}}(\boldsymbol{\mu}^{(g)}|\mathbf{W}^{(g)})$, and a group sparsity constraint $\sum_t \Omega(\mathbf{h}_t)$. The first term is a generalized KL divergence used by the basic un-supervised NMF algorithm. The second term is a constraint which allocates sub-groups of bases $\mathbf{W}^{(g)}$ to g th acoustic atoms characterized by the centroid $\boldsymbol{\mu}^{(g)}$. The strength of constraint is controlled by η. The third term represents group sparsity constraint at time t controlled by λ, where

$$\Omega(\mathbf{h}_t) = \sum_g \log(\epsilon + \|\mathbf{h}_t^{(g)}\|_1) \quad (4)$$

is used in prior arts [16, 20] to turn off activation of the irrelevant acoustic atoms.

To minimize the cost function in (2), the following update rules are iteratively applied:

$$\mathbf{W}^{(g)} \leftarrow \mathbf{W}^{(g)} \odot \left\{ \left(\frac{\mathbf{V}}{\boldsymbol{\Lambda}} \right) \mathbf{H}^\top + \eta \frac{\boldsymbol{\mu}^{(g)}}{\mathbf{W}^{(g)}} \right\} \Big/ \left\{ \mathbf{1} \left(\mathbf{H}^\top + \eta \right) \right\}, \quad (5)$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left\{ \mathbf{W}^\top \left(\frac{\mathbf{V}}{\boldsymbol{\Lambda}} \right) \right\} \Big/ \left\{ \mathbf{W}^\top \mathbf{1} \right\}, \quad (6)$$

$$\mathbf{h}_t^{(g)} \leftarrow \mathbf{h}_t^{(g)} \frac{1}{1 + \lambda \Big/ \left\{ \epsilon + ||\mathbf{h}_t^{(g)}||_1 \right\}} \quad (7)$$

where $\mathbf{1}$ is a matrix with all elements equal to 1 and with a dimension of $\mathbf{V}$. $\mathbf{A} \odot \mathbf{B}$ represents element wise multiplication, $\mathbf{A}/\mathbf{B}$ and $\frac{\mathbf{A}}{\mathbf{B}}$ represent element wise division. The procedure of dictionary learning is shown in Algorithm 1.

2.2. Classifier training

In classifier training, an activation matrix $\mathbf{H}_{(i)}$ is extracted from the corresponding training spectrogram $\mathbf{V}_{(i)}$ by supervised NMF with MLD $\mathbf{W}$ and a classifier is trained using the activation matrices where $i \in \{1, ..., I\}$ represents an event-class index. In supervised NMF, $\mathbf{V}_{(i)}$ is approximated by a product of $\mathbf{W}$ and $\mathbf{H}_{(i)}$,

$$\mathbf{V}_{(i)} \sim \mathbf{WH}_{(i)}. \quad (8)$$

Algorithm 1 Dictionary learning for MLD

1: INPUT: $\mathbf{V}$
2: Obtain $\mathbf{W}_0$ by a basic NMF
3: Obtain $\boldsymbol{\mu}^{(g)}$ by using K-means to $\mathbf{W}_0$
4: Initialize $\mathbf{W}$ and $\mathbf{H}$ with random values.
5: **repeat**
6: Update $\mathbf{W}$ using (5).
7: Update $\mathbf{H}$ using (6) and (7).
8: **until** Convergence
9: OUTPUT: $\mathbf{W}$

For a given $\mathbf{W}$ by dictionary learning, $\mathbf{H}_{(i)}$ is updated using (6) and the group sparsity constraint in (7). The procedure is shown in Algorithm 2.

Once $\mathbf{H}_{(i)}$ has been obtained, column vectors $\mathbf{h}_{t(i)}$ of $\mathbf{H}_{(i)}$ at each time frame t are used as feature vectors to train the classifier. Simple linear support vector machine(SVM) [21] is used for classifier. Multi-class SVM is trained based on the one-against-all approach.

Algorithm 2 Feature extraction with supervised NMF

1: INPUT: $\mathbf{V}_{(i)}$ and $\mathbf{W}$
2: Initialize $\mathbf{H}_{(i)}$ with random values.
3: **repeat**
4: Update $\mathbf{H}_{(i)}$ using (6) and (7). with fixed $\mathbf{W}$
5: **until** Convergence
6: OUTPUT: $\mathbf{H}_{(i)}$

2.3. Event classification

In event classification, the proposed method extracts an activation matrix from a test spectrogram using semi-supervised NMF with MLD. A noise dictionary is learned concurrently with extracing the activation matrix. Unknown spectra included in a test spectrogram is expressed by the noise dictionary and the corresponding activation matrix, so that an activation matrix of acoustic atoms are extracted properly.

Let $\mathbf{V}_{(*)}$ and $\mathbf{W}_n \in \mathcal{R}_+^{F\times K_n}$ denote the test spectrogram and the noise dictionary, respectively, where K_n is the number of bases in the noise dictionary. $\mathbf{H}_{(*)}$ and $\mathbf{H}_n$ denote activation matrices of MLD and $\mathbf{W}_n$, respectively. The relationship among these matrices is described as in the following approximation:

$$\mathbf{V}_{(*)} \sim \mathbf{\Lambda}_{(*)} = [\mathbf{W}, \mathbf{W}_n] \begin{bmatrix} \mathbf{H}_{(*)} \\ \mathbf{H}_n \end{bmatrix}. \tag{9}$$

In semi-supervised NMF, $\mathbf{H}_{(*)}$, $\mathbf{H}_n$ and $\mathbf{W}_n$ are updated to minimize a generalized KL divergence $\mathcal{D}_{\mathcal{KL}}\left(\mathbf{V}_{(*)}|\mathbf{\Lambda}_{(*)}\right)$, applying an update rule for the activation matrix in (6), a group sparsity constraint in (7) and the following update rule for $\mathbf{W}_n$:

$$\mathbf{W}_n \leftarrow \mathbf{W}_n \odot \left\{ \left(\frac{\mathbf{V}_{(*)}}{\mathbf{\Lambda}_{(*)}} \right) \mathbf{H}_n^\top \right\} \Big/ \left\{ \mathbf{1}\mathbf{H}_n^\top \right\}. \tag{10}$$

The procedure is shown in Algorithm 3.

Figure 2 is a simple illustration of the relationship among training/test spectrum, MLD, and the noise dictionary. The relationship is explained as data points on the 3-dimentional simplex [22, 23]. ◯ and ★ represent training and test spectrum, respectively, ◊ and △ represent bases and centroids of MLD, respectively. In dictionary learning, sub-groups of bases ◊ in MLD are learned to span convex hulls enclosing training spectra ◯. In event classification, the noise dictionary is learned from unknown test spectra ★ lying outside the convex hulls which is indicated with the shaded area. Therefore unknown spectra included in the test spectrogram are expressed by the noise dictionary and MLD can extract a proper activation matrix of acoustic atoms.

After extracting $\mathbf{H}_{(*)}$, the classifier receives $\mathbf{H}_{(*)}$ as a feature and outputs a $T \times I$ binary classification-result matrix $\mathbf{R}$, where I represents the number of event classes for classification. A binary column vector of $\mathbf{R}$ per frame corresponds to the presence of each event class. When a column of $\mathbf{R}$ contains two non-zero elements for example, there are two detected events in that frame. A non-zero and a zero column vector stand for event-detected and event-undetected status, respectively.

Algorithm 3 Feature extraction with semi-supervised NMF

1: INPUT: $\mathbf{V}_{(*)}$ and $\mathbf{W}$
2: Initialize $\mathbf{W}_n$, $\mathbf{H}_{(*)}$ and $\mathbf{H}_n$ with random values.
3: **repeat**
4: Update $\mathbf{W}_n$ using (10)
5: Update $\mathbf{H}_{(*)}$ and $\mathbf{H}_n$ using (6) and (7).
6: **until** Convergence
7: OUTPUT: $\mathbf{H}_{(*)}$

3. ADDITIONAL PROCESSING SPECIFIC TO EVALUATION

DCASE 2016 task 2 Dataset is used for evaluating the proposed method. The Dataset includes 11 event classes, which are typically found in the office and shown on the left side of Figure 4. The task 2 has two types of dataset; Training Dataset used for learning MLD and training SVM classifiers, and Development Dataset used for classification.

Training Dataset consists of 20 noise-free files for each event class totaling 220 files. Development Dataset includes 18 files to cover 6 event occurrence patterns and three SNRs, namely, -6, 0, and 6 dB, each of which contains all 11 event classes. Development Dataset also has an annotation file for each sound data file to evaluate classification result.

Because DCASE 2016 task 2 Dataset is used for evaluation with the annotation file and classification results, each classification result needs to be expressed in the format of the annotation file, which is defined as columns of event class name, onset time, and offset time. The classification-result matrix $\mathbf{R}$ is applied a median filter in a row-wise manner. Columns of $\mathbf{R}$ are further replaced with zero column vectors when the corresponding frame is determined as silent by an integrated spectral intensity (ISI) or as a gap shorter than 0.1 second (10 frames). Values of F-measure are calculated by sed_eval tools [24] for evaluation on segment based metrics over 1.0 second for each SNR and each event.

4. EVALUATION AND DISCUSSION

Table 1 shows a parameter setting used in the evaluation. For generating spectrograms from sound files, a variable q transform (VQT) [25] was used. VQT spectrograms were extracted for all files of

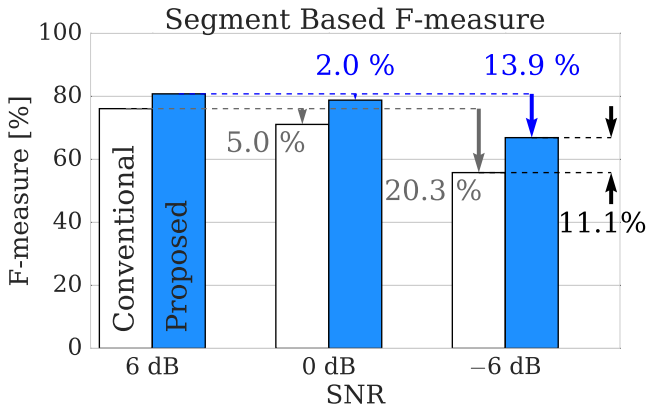

Figure 3: Evaluation results for all event classes at three different SNRs using DCASE2016 task 2 Dataset.

Table 1: Parameter setting for the evaluation.

Parameter	value	
Sampling rate	44.1	kHz
F_{min} for VQT	27.5	Hz
Number of bins per octave for VQT	60	
γ for VQT	30.0	
Number of basis for MLD	46	
Number of group basis for MLD	4	

DCASE task 2 Dataset. MLD was learned from the obtained VQT spectrograms. The number of bases in the noise dictionary for semi-supervised NMF was set to the one with the best performance for each event class.

Figure 3 compares F-measure values calculated by the conventional AED with supervised NMF and the proposed AED with semi-supervised NMF for different SNRs. The F-measures by the proposed method are 4.7%, 7.7%, and 11.1% higher than those by the conventional method at SNRs of 6, 0, and −6 dB, respectively. The degradation of F-measure from 6 to 0 dB is 2.0% and that from 6 to −6 dB is 13.9% for the proposed method. These values are smaller than those for the conventional method.

Conventionally, the input spectrogram including the noise is modeled by fixed MLD, which is learned without noise, and the activation matrix of MLD, so that the activation matrix of MLD includes errors. The proposed method dedicates both a noise dictionary and its activation matrix to the noise. Because noise spectra are better modeled by the noise dictionary learned upon classification and its activation matrix, the proposed method provides a higher F-measure values than conventional method at each SNR, when known acoustic atoms in the learning data are contaminated by noise in event classification. Therefore, the proposed method is robust to the noise.

Results for each event class are compared in Figure 4. It shows big improvement for cough and page turn. Especially, the F-measure of page turn is improved by 24.4%. The F-measure for clear throat, keyboard, keys, laughter, phone, and speech show small improvement. Door slam, drawer, and knock did not improve at all. The proposed method generally provides better results than the conventional method for each event class, because the

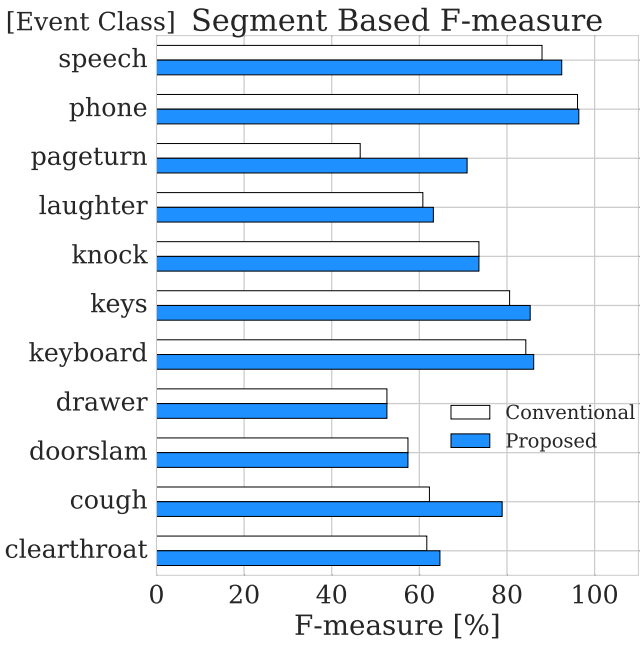

Figure 4: Evaluation results for each acoustic event of DCASE2016 task 2 Dataset.

conventional method is a special case of the proposed method with no noise dictionary and no noise activation matrix. The effect of the proposed method changes according to similarities between an event-class spectrum and an unknown noise spectrum. It seems that an event class with big improvement by the proposed method has MLD that can be easily activated by the noise spectrum. The proposed method reduces such erroneous activation with a help of the noise dictionary. In contrast, when the spectrum of an event class are clearly different from the noise spectrum, the proposed method is not as effective as for the similar spectrum case. Further investigation is left for future study.

5. CONCLUSIONS

An acoustic event detection (AED) method using semi-supervised non-negative matrix factorization (NMF) with a mixture of local dictionaries (MLD) has been proposed. The proposed method has newly introduced a noise dictionary and a noise activation matrix both dedicated to unknown acoustic atoms which are not included in MLD. Because unknown acoustic atoms are better modeled by the new noise dictionary learned upon classification and the new activation matrix, the proposed method provides a higher classification performance for event classes modeled by MLD when a signal to be classified is contaminated by unknown acoustic atoms. Evaluation results using DCASE2016 task 2 Dataset have shown that F-measure by the proposed method with semi-supervised NMF has been improved by as much as 11.1% compared to that by the conventional method with supervised NMF.

6. REFERENCES

[1] D. Hollosi, J. Schröder, S. Goetze, and J. -E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," in *2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010)*. IEEE, 2010, pp. 1–5.

[2] J. Schröeder, S. Wabnik, P. W. Van Hengel, and S. Goetze, "Detection and classification of acoustic events for in-home care," in *Ambient Assisted Living*. Springer, 2011, pp. 181–195.

[3] S. Chu, S. Narayanan, and C. -C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[4] G. Roma, J. Janer, S. Kersten, M. Schirosa, P. Herrera, and X. Serra, "Ecological acoustics perspective for content-based retrieval of environmental sounds," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, p. 1, 2010.

[5] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 4, no. 2, p. 11, 2008.

[6] 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), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3163–3167.

[7] 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), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6255–6259.

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

[9] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6460–6464.

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

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

[12] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[13] E. Benetos, G. Lafay, M. Lagrange, and M. Plumbley, "Detection of overlapping acoustic events using a temporally-constrained probabilistic model," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6450–6454.

[14] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, *et al.*, "An exemplar-based nmf approach to audio event detection," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[15] T. Komatsu, Y. Senda, and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2259–2263.

[16] 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.

[17] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[18] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 414–421.

[19] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 556–562.

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

[21] R. -E. Fan, K. -W. Chang, C. -J. Hsieh, X. -R. Wang, and C. -J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[22] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*, 2004, pp. 1141–1148.

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

[24] DCASE2016, "Detection and classification of acoustic scenes and events 2016," http://www.cs.tut.fi/sgn/arg/dcase2016/.

[25] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society Conference, 53rd International Conference on*, 2014.