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 the 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 the 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 promising solutions [10, 11, 12, 13]. For 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 atoms. The MLD is learned with constrained 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, the 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 the 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 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. Figure 2 shows a relationship among training/test spectrum, the MLD, and the noise dictionary.

MLD consists of $G$ sub-groups of bases which model acoustic atoms $W = [W^{(1)}, ..., W^{(G)}]$. A basis matrix $W^{(g)} \in \mathbb{R}^{F_{g} \times K_{g}}$ consists of $K_{g}$ basis vectors where $\mathbb{R}^{F_{g} \times K_{g}}$, $F_{g}$, 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, respectively.

To determine acoustic atoms, an initial basic matrix $W_0$ is first extracted from the entire training data spectrogram $V \in \mathbb{R}^{F \times T}$ with the basic un-supervised NMF where $T$ denotes its number of time frames. K-means clustering is then applied to bases in $W_0$ to select $G$ centroids $\mu^{(g)}$ which represent centroids of acoustic atoms.

NMF is again applied to $V$ with the centroids $\mu^{(g)}$ and the following cost function $D(V|\Lambda)$:

$$D(V|\Lambda) = D_{KL}(V|\Lambda) + \lambda \sum_{g} \Omega(h_{t}^{g}) + \eta \sum_{g} \sum_{n} \mu^{(g)}_{n} \left| h_{t}^{(g)n} \right|_{1}$$  

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

$$H = [h_{1}, ..., h_{t}, ..., h_{T}],$$  

$$h_{t}^{T} = [h_{t}^{(1)} \top, ..., h_{t}^{(g)} \top, ..., h_{t}^{(G)} \top],$$  

where $[\cdot]^{T}$ denotes a matrix transpose.

Cost function in (2) consists of three terms; a generalized Kullback-Leibler(KL) divergence $D_{KL}(V|\Lambda)$ between $V$ and $\Lambda$, a constraint $\sum_{g} \Omega(h_{t}^{g})$, and a group sparsity constraint $\sum_{g} \mu^{(g)}_{n}$.

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 $W^{(g)}$ to $g$ the acoustic atoms characterized by the centroid $\mu^{(g)}$. The strength of constraint is controlled by $\eta$. The third term represents group sparsity constraint at time $t$ controlled by $\lambda$, where

$$\Omega(h_{t}) = \sum_{g} \log(\epsilon + \left| h_{t}^{(g)} \right|_{1})$$

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:

$$W^{(g)} \leftarrow W^{(g)} \odot \left\{ \frac{(V\Lambda)^{T} + \eta \mu^{(g)}_{n}}{\left| W^{(g)} \right|^{T} + 1} \right\},$$

$$H \leftarrow H \odot \left\{ \frac{W^{T}(V\Lambda)}{\left| W^{T} \right| + 1} \right\},$$

$$h_{t}^{(g)n} \leftarrow \frac{1}{1 + \lambda \left\{ \epsilon + \left| h_{t}^{(g)n} \right|_{1} \right\}} h_{t}^{(g)n}$$

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

**Algorithm 1** Dictionary learning for MLD

1: **INPUT:** $V$
2: Obtain $W_0$ by a basic NMF
3: Obtain $\mu^{(g)}$ by using K-means to $W_0$
4: Initialize $W$ and $H$ with random values.
5: **repeat**
6: Update $W$ using (5).
7: Update $H$ using (6) and (7).
8: **until** Convergence
9: **OUTPUT:** $W$
2.2. Classifier training

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

$$V_{(i)} \approx WH_{(i)}.$$  

(8)

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

Once $H_{(i)}$ has been obtained, column vectors $H_{(i)}$ of $H_{(i)}$ at each time frame $t$ are used as feature vectors to train the classifier. Simple linear support vector machine (SVM) [21] is used for each time frame.

Algorithm 2 Feature extraction with supervised NMF

1: INPUT: $V_{(i)}$ and $W$
2: Initialize $H_{(i)}$ with random values.
3: repeat
4: Update $H_{(i)}$ using (6) and (7), with fixed $W$
5: until Convergence
6: OUTPUT: $H_{(i)}$

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 extracting the activation matrix. Unknown spectra included in the test spectrogram are expressed by 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 $H_{(s)}$, the classifier receives $H_{(s)}$ as a feature and outputs a $T \times I$ binary classification-result matrix $R$, where $I$ represents the number of event classes for classification. A binary column vector of $R$ per frame corresponds to the presence of each event class. When a column of $R$ 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: $V_{(s)}$ and $W$
2: Initialize $W_n$, $H_{(s)}$ and $H_n$ with random values.
3: repeat
4: Update $W_n$ using (10)
5: Update $H_{(s)}$ and $H_n$, using (6) and (7).
6: until Convergence
7: OUTPUT: $H_{(s)}$

3. ADDITIONAL PROCESSING SPECIFIC TO EVALUATION

DCASE 2016 task 2 Dataset is used for evaluating the proposed method. The Dataset includes 11 sound 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 generating MLD and training SVM classifiers, and Development Dataset used for classification.

Training Dataset consists of 20 noise-free files for each sound 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 sound 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 sound class name, onset time, and offset time. The classification-result matrix $R$ is applied a median filter in a row-wise manner. Columns of $R$ 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 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 DCASE task 2 Dataset. MLD was generated 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[18] 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
Evaluation results using DCASE2016 task 2 Dataset. Table 1: Parameter setting for the evaluation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
<td>44.1 kHz</td>
</tr>
<tr>
<td>$f_{\text{min}}$ for VQT</td>
<td>27.5 Hz</td>
</tr>
<tr>
<td>Number of bins per octave for VQT</td>
<td>60</td>
</tr>
<tr>
<td>$\gamma$ for VQT</td>
<td>30.0</td>
</tr>
<tr>
<td>Number of basis for MLD</td>
<td>46</td>
</tr>
<tr>
<td>Number of group basis for MLD</td>
<td>4</td>
</tr>
</tbody>
</table>

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%. 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 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.

6. CONCLUSIONS

An acoustic event detection (AED) method using semi-supervised non-negative matrix factorization (NMF) with 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 the learning data. Because unknown acoustic atoms are better modeled by the fix noise dictionary learned upon classification and the new activation matrix, the proposed method provides a higher classification capability 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.
7. REFERENCES


