OVERLAPPING SOUND EVENT DETECTION WITH SUPERVISED NONNEGATIVE MATRIX FACTORIZATION

Victor Bisot, Slim Essid, Gaël Richard

LTIC, CNRS, Télécom ParisTech, Université Paris - Saclay, F-75013, Paris, France

<firstname>.<lastname>@telecom-paristech.fr

ABSTRACT

In this paper we propose a supervised Nonnegative Matrix Factorization (NMF) model for overlapping sound event detection in real life audio. We start by highlighting the usefulness of non-euclidean NMF to learn representations for detecting and classifying acoustic events in a multi-label setting. Then, we propose to learn a classifier and the NMF decomposition in a joint optimization problem. This is done with a general $\beta$-divergence version of the nonnegative task-driven dictionary learning model. An experimental evaluation is performed on the development set of the DCASE 2016 task3 challenge. The proposed supervised NMF-based system improves performance over the baseline and the submitted systems.

Index Terms— Acoustic Event Detection, Nonnegative Matrix Factorization, Supervised Feature learning

1. INTRODUCTION

Acoustic event detection (AED) is the task of transcribing an audio recording into a symbolic description representing the sound events occurring in an acoustic scene. AED requires a system to be able to classify the events present in the recording as well as their position in time. The interest for AED has kept increasing in the last few years in part due to the release of new datasets along with the organization of international evaluation campaigns including CLEAR 2007 and the DCASE 2013 and 2016 challenges [1, 2]. Enabling devices to be aware of the different acoustic events occurring in their surroundings has a wide variety of potential applications such as surveillance [3], health monitoring [4] or multimedia indexing [5].

The AED task can be separated into two different sub-problems depending on whether the events to be detected can overlap in time. The first one is referred to as monophonic event detection, where only one event label can be present at any given time. Therefore, this setting is closer to a standard multi-class classification problem. It has often been addressed using speech inspired techniques. Indeed, a common approach is to represent local frames by Mel-Frequency Cepstral Coefficient (MFCC) features modeled by Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) emission probabilities for detection [6]. Another popular approach is to consider standard classifiers such as support vector machines [7] trained on sliding windows of the signal.

The second problem, much more challenging, is polyphonic event detection. In this case, the different events in the scene can overlap in time, turning AED into a multi-label classification problem. A first possible approach is to train as many binary event/background classifiers as there are possible event labels [2]. The other, more common strategy, is to threshold the outputs of regression or generative probabilistic models, allowing multiple labels to be detected at a given time. A few works considered the use of Nonnegative Matrix Factorization (NMF), where spectral templates are learned on isolated events. Then, detection is performed by applying a threshold to the activation matrix obtained from decomposition of the test data. For example, this has been done using a PLCA model constrained with a HMM [8] or by decomposing the data matrix and a label matrix in a joint matrix factorization problem [9]. More recently, especially with the organization of the DCASE 2016 challenge, the popularity of Neural Network-based systems for AED has strongly increased. This trend has started prior to the challenge with the introduction of Recurrent Neural Networks (RNN) for the task [10] offering improvements over the standard approaches on a private dataset. Thereafter, most of the challenge submissions were comprised of a variety of different deep neural networks models including Recurrent Neural Networks (RNN) [11], Gated RNNs [12] and Convolutional NNs [13]. Interestingly, only one of them managed to outperform the MFCC/GMM baseline on the final evaluation set [11].

In this paper, we further study the usefulness of NMF for overlapping AED in real world audio by showing the potential of both unsupervised and supervised NMF. Contrary to other NMF-based methods for AED, we do not have access to the isolated events during the training. In fact we work in the context of the DCASE 2016 challenge where only annotated real life recordings are available for training. Here we use NMF as a feature learning technique, where a dictionary of spectral templates is learned from the training data before classifying separately the projections on that dictionary. We deal with the multi-label problem by training a multinomial logistic regression classifier only with the frames that do not contain overlapping events, and then threshold the output probabilities during the detection stage. We then study the usefulness of a supervised NMF model referred to as Nonnegative task driven dictionary learning (TD-NMF) originally introduced in [14, 15]. In TD-NMF, the classifiers and dictionaries are learned in a bi-level optimization problem in order to obtain more discriminant dictionaries. This is a first attempt at using the TD-NMF with the $\beta$-divergence for a multi-label classification problems. Our NMF-based systems are evaluated on the development set of the 2016 DCASE challenge. We show that both the supervised and unsupervised versions yield performance which is competitive with the best neural network-based systems submitted to the challenge. Finally, we also discuss the potential and benefits of non-Euclidean TD-NMF for polyphonic AED. Its usefulness for detecting overlapping events is highlighted on a novel evaluation paradigm, by scoring only on parts of the audio containing overlapping events in the annotation.

The paper is organized as follows. The problem and the general NMF approach are introduced in Section 2. The TD-NMF model is
2. ACoustIC EVENT DETECTION SYSTEM

2.1. Input representation

The NMF stage takes as input the time frequency representation \( V \in \mathbb{R}^{F \times N} \) of the audio signals, where \( F \) is the number of frequency bands and \( N \) is the number of time frames. The nonnegative matrix \( V \) is build by concatenating the Mel-Spectrum extracted from each recording in a training set. While many AED works rely on MFCC as their input representation, it is also common to use perceptually motivated time-frequency representation such as Mel-spectrograms or Constant Q-transforms.

2.2. Unsupervised nonnegative matrix factorization

NMF is a well known technique to decompose nonnegative data into nonnegative dictionary elements [16]. The goal of NMF is to find a decomposition that approximates the data matrix \( V \) such as:

\[
V \approx WH, \quad (1)
\]

with \( W \in \mathbb{R}^{F \times K} \) and \( H \in \mathbb{R}^{K \times N} \). Many problems benefit from the nonnegativity of the decomposition to learn better representations of the data, especially in the audio processing field. In fact, most of the time-frequency representations for audio signals contain only nonnegative coefficients. For multi-source environments like in AED, the nonnegative constraints allow for interpreting the time-frequency representation as a sum of different nonnegative sound objects, corresponding to the different events occurring in the scene. Given a separable divergence \( D_\beta \), NMF is obtained by solving the following optimization problem:

\[
\min_{W,H} D_\beta(V|WH) \text{ s.t. } W, H \geq 0. \quad (2)
\]

In this work \( D_\beta \) represents the \( \beta \)-divergence [17]. The particular cases of interest for the \( \beta \)-divergence are the Euclidean distance (\( \beta = 2 \)), Kullback-Leibler (\( \beta = 1 \)) and Itakura-Saito (\( \beta = 0 \)). One property of the \( \beta \) parameter is that it influences the divergence’s invariance to a scaling factor. Many audio signal processing applications often benefit from the non-Euclidean cases when performing NMF.

2.3. Classification stage

After jointly learning the dictionary and the projections on the training data, we project the test data on the learned dictionary. The projections are then used as features for classification. The classifier is a regularized linear logistic regression in its multinomial formulation. Logistic regression has the advantage of outputting the class probabilities for each data point. However, being a multi-class classifier, it can not directly deal with the multi-label classification problem. In order to overcome this limitation, we simply drop the frames associated with multiple labels in the training data. We then add a label corresponding to the background, when no events are labeled. This strategy allows us to train the classifier in a standard multi-class fashion.

Then, in order to allow the system to predict multiple labels during the test detection stage, we threshold the class probabilities for each test frame. Therefore, overlapping events can be detected as long as their probability in the given frame is above the fixed threshold.

3. SUPERVISED NONNEGATIVE MATRIX FACTORIZATION

The motivation behind using supervised NMF is to merge the NMF and the classification stage in a common bi-level optimization problem. The goal is then to learn discriminative nonnegative dictionaries of spectral templates with the objective of minimizing the classification cost. We start from a variant of the original Task-driven dictionary learning model [14], which proposes a TD-NMF algorithm to jointly learn an NMF with a multinomial logistic regression classifier for acoustic scene classification [18]. Then, as the TD-NMF algorithm in [18] is limited to the Euclidean case, we extend it to include the possibility of using the Kullback-Leibler reconstruction cost. These changes are inspired from a previous work [15], which proposed a first application of nonnegative task-driven dictionary learning, by linking \( \beta \)-divergence NMF to a speech enhancement criteria. Finally, in our case, the TD-NMF model jointly learns a multinomial logistic regression and a NMF with a \( \beta \)-divergence reconstruction cost.

The TD-NMF model first considers the optimal projections \( h^*(v, W) \) of the data point \( v \) on the dictionary \( W \). The projections are defined as solutions of the nonnegative projection with \( \ell_1 \) and \( \ell_2 \)-norm penalties expressed as in an elastic-net fashion:

\[
h^*(v, W) = \min_{h \in \mathbb{R}_+^K} D_\beta(V|Wh) + \lambda_1 \|h\|_1 + \lambda_2 \|h\|_2^2; \quad (3)
\]

where \( \lambda_1 \) and \( \lambda_2 \) are nonnegative regularization parameters. Let each data frame \( v \) be associated with a label \( y \) in a fixed set of labels \( \mathcal{Y} \). We then define the classification loss to be a multinomial logistic loss \( l_s(y, A, h^*(v, W)) \), a function of the optimal projection \( h^*(v, W) \), where \( A \subseteq \mathcal{A} \) are the parameters of the classifier. The TD-NMF problem is now expressed as a joint minimization of the expected classification cost over \( W \) and \( A \):

\[
\min_{W,A} \min_{h \in \mathbb{R}_+^K} f(W, A) + \nu \|A\|_2^2. \quad (4)
\]

where

\[
f(W, A) = E_{y,A}[l_s(y, A, h^*(v, W))]. \quad (5)
\]

Here, \( W \) is defined as the set of nonnegative dictionaries containing unit \( l_2 \)-norm basis vectors and \( \nu \) is a regularization parameter on the classifier parameters to prevent over-fitting. The TD-NMF model in equation (4) is optimized with a stochastic gradient descent algorithm as described in [18]. For the general \( \beta \) case, the expression of the gradient of \( f(W, A) \) with respect to the dictionary \( W \) is provided in [15]. It is important to note that these expressions only hold for \( \beta \geq 1 \) as they rely on the convexity of equation (3). Therefore, we will only be considering the Euclidean (\( \beta = 2 \)) and Kullback-Leibler (\( \beta = 1 \)) versions of TD-NMF.

Once the model has been trained in the standard multi-class setting, the multi-label prediction step is similar to one described in Section 2.3. The data is projected on the learned discriminative dictionary before thresholding the output class probabilities from the model classifier.
4. EXPERIMENTAL EVALUATION

4.1. Dataset

Our proposed methods are evaluated on the development dataset of the DCASE 2016 challenge for sound event detection in real life audio (corresponding to the task 3) [2]. To our best knowledge, it is the only publicly available non-synthetic polyphonic acoustic event detection dataset proposing different environments. Each audio recording in the dataset is associated with manually annotated onset and offset event timings. The recordings were done in two different acoustic environments: *home* (indoors) and *residential area* (outdoors). There are 10 to 12 recordings of 3 to 5 minutes duration for both environments, which are split into 4 cross-validation folds where each recording is used once in the test data. The list of event labels as well as their number of occurrences are outlined in Table 1. The results of the evaluation campaign later confirmed it to be a particularly challenging dataset. This can be attributed to the subjectivity of the manual annotations as well as the lack of training data, resulting in some events having a low amount of training examples.

4.2. Performance metrics

In a first stage we use the same performance metrics to rank the submissions as in the DCASE challenge. The first one is the segment-based F1 score. It is computed as the harmonic mean between precision and recall based on the total amount of false negatives, true positives and false positives in one second segments. The second metric is the segment-based acoustic event error rate (ER). The ER is calculated by adding the number of substitutions, insertions and deletions in each 1-second segment before dividing it by the total number of events. The ER and F1 scores are computed over the entire test set by evaluating all 4 test folds at once. We refer the reader to [19] for more details and explanations about these metrics.

4.3. Experimental setup

The Mel-spectrum were extracted with the YAAFE toolbox [20] after rescaling the signals in [−1, 1]. They were computed using 40 Mel-bank filters on 40 ms frames with 50% overlap. The training data matrix was scaled in order to have unit variance, the same scaling factors were then applied to the test data.

The unsupervised NMF was applied in the \( \ell_1 \) sparse formulation [21] and optimized using multiplicative update rules as described in [22]. The classifier was trained using the scikit-learn [23] implementation of the multinomial logistic regression using the L-BFGS solver and the \( \nu \) regularization parameter was fixed to \( \nu = 10 \) after testing.

The TD-NMF model is initialized with the \( W \) and \( A \) learned from the unsupervised NMF system. For the Euclidean case (\( \beta = 2 \)), the optimal projections in equation (3) were obtained using the *lasso* function from the *spams* toolbox [24]. Whereas for other values of \( \beta \), equation (3) was solved using multiplicative update rules. The model was trained over \( I = 6 \) iterations (epochs) with a 0.001 initial gradient step for the projected gradient dictionary update. The decaying of the gradient step over iterations follows the same heuristic as suggested in [14]. The \( \ell_1 \) and \( \ell_2 \) regularization parameters were set to \( \lambda_1 = 0.5 \) and \( \lambda_2 = 0 \) for both the NMF and TD-NMF. Modifying those values did not provide any significant modifications on the performance for both environments. After learning discriminative dictionaries with TD-NMF, the data was fully reprojected on that dictionary and the classifier was updated until convergence. Finally, the threshold for detection on the output probabilities was set to 0.3 for the *home* environment and to 0.35 for *residential area*. These values were chosen after testing as they provided a good compromise between precision and recall. This difference is due to the probabilities being more spread out for *home* as it contains for possible outputs. The predictions were then filtered with a median filter long of 7 time frames in order to discard outliers and shorter events.

4.4. Results with the challenge metrics

The results of both the NMF and TD-NMF using the standard metrics for the DCASE dataset are reported in Table 2. First, for every configuration presented, the TD-NMF model outperforms the unsupervised NMF. This shows that the supervised model is able to learn more discriminative dictionaries representing the events in the dataset. As the results slightly increase when augmenting the dictionary size for NMF, TD-NMF performs just as well with smaller dictionaries, keeping only the relevant information to minimize the classification cost. As for the \( \beta \) parameter, decomposing the data with the Kullback-Leibler divergence (\( \beta = 1 \)) slightly increases the performance over using the Euclidean distance for the unsupervised
In this paper we presented a supervised NMF-based approach to polyphonic AED. The TD-NMF approach has the advantage of jointly learning a dictionary and a multinomial logistic regression while being able to output overlapping labels. This TD-NMF formulation also allows the choice of the $\beta$-divergence in the NMF problem. An experimental evaluation was done on the development set of the DCASE 2016 challenge for AED in real life audio. The results show that both unsupervised and supervised NMF systems can compete with DNN-based methods. TD-NMF showed improvements over the best systems on this dataset. Finally, the potential of the Kullback-Leibler TD-NMF for highly polyphonic AED was highlighted by an evaluation performed on the segments containing overlapping events. For future works, the focus will be on incorporating temporal modeling to the model in order to increase the precision of the detection.

5. CONCLUSION

Table 3: Error rate and F1 score on 1 second segments for NMF and TD-NMF compared to the best methods on the DCASE development dataset. The average results over both environments are reported.

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Average ER</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2] GMM</td>
<td>MFCC</td>
<td>91</td>
<td>23.7</td>
</tr>
<tr>
<td>[25] Random Forests</td>
<td>MFCC</td>
<td>76</td>
<td>38.5</td>
</tr>
<tr>
<td>[12] GRNN</td>
<td>Spectrogram</td>
<td>73</td>
<td>47.6</td>
</tr>
<tr>
<td>[26] RNN</td>
<td>Mel energy</td>
<td>81.5</td>
<td>49.8</td>
</tr>
</tbody>
</table>

Table 4: Error rate and F1 score on 100 ms segments for TD-NMF in function of the dictionary size $K$ and the divergence $\beta$. The all segments denotes the case where all segments are used for evaluation and the with overlap only keeps segments containing overlap. The best scores are highlighted in bold text.

<table>
<thead>
<tr>
<th>Method</th>
<th>$K$</th>
<th>$\beta$</th>
<th>All segments</th>
<th>ER</th>
<th>F1</th>
<th>with overlap</th>
<th>ER</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TD-NMF</td>
<td>8</td>
<td>2</td>
<td>61</td>
<td>56</td>
<td>75</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TD-NMF</td>
<td>8</td>
<td>1</td>
<td>59</td>
<td>58</td>
<td>71</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TD-NMF</td>
<td>16</td>
<td>2</td>
<td>58</td>
<td>59</td>
<td>75</td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TD-NMF</td>
<td>16</td>
<td>1</td>
<td>60</td>
<td>57</td>
<td>70</td>
<td>43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6. REFERENCES


