ORTHOGONALITY-REGULARIZED MASKED NMF FOR LEARNING ON WEAKLY LABELED AUDIO DATA

Iwona Sobieraj, Lucas Rencker, Mark D. Plumbley

University of Surrey
Centre for Vision Speech and Signal Processing
Guildford, Surrey GU2 7XH, United Kingdom

ABSTRACT

Non-negative Matrix Factorization (NMF) is a well established tool for audio analysis. However, it is not well suited for learning on weakly labeled data, i.e. data where the exact timestamp of the sound of interest is not known. In this paper we propose a novel extension to NMF, that allows it to extract meaningful representations from weakly labeled audio data. Recently, a constraint on the activation matrix was proposed to adapt for learning on weak labels. To further improve the method we propose to add an orthogonality regularizer of the dictionary in the cost function of NMF. In that way we obtain appropriate dictionaries for the sounds of interest and background sounds from weakly labeled data. We demonstrate that the proposed Orthogonality-Regularized Masked NMF (ORM-NMF) can be used for Audio Event Detection of rare events and evaluate the method on the development data from Task2 of DCASE2017 Challenge.

Index Terms— Non-negative Matrix Factorization, weakly labeled data, Acoustic Event Detection

1. INTRODUCTION

Non-negative Matrix Factorization (NMF) is a popular tool for discovering structure in a variety of signals. For many years it has been widely used for analysis of musical audio and more recently, of environmental sounds. Analysis of environmental sounds have recently received a lot of attention in the research community due to its vast number of applications, ranging from audio content analysis, human activity monitoring, to surveillance and bioacoustic monitoring. Among others, methods based on NMF have been successfully applied to several tasks of environmental audio analysis such as audio scene classification [1], rare event detection [2] or real life audio event detection [3]. NMF methods offer parsimonious models with significantly fewer parameters than, for instance, Deep Neural Networks (DNNs). Hence further investigating NMF is an interesting direction in the environmental sound research.

NMF models the spectrogram $\mathbf{V}$ of a sound signal as a product of a dictionary of spectral bases $\mathbf{W}$ and a corresponding activation matrix $\mathbf{H}$. The key strategy for NMF to efficiently model the sounds of interest is to express them and the background sounds by different sets of bases. It is easily achieved when we have access to isolated recordings of sound of interest [2] or well annotated data, where the timestamps of the sounds of interest are known [3]. In that case, we can extract the set of bases for the sound of interest using the isolated/annotated recordings and another set of bases for the background sounds using the recordings of the noise. However, in the real world scenario it is often easier to gather *weakly labeled* data, that is, data in which we do not have exact information of when the interesting sound occurs, but just a tag of which sounds are present in a given audio excerpt. It implies that we do not have access to the clean recordings of the sound of interest: the training data contain parts with background/noise only and parts with background and the target sound. Therefore, the task of expressing the sound of interest and the background with different bases becomes difficult. In [4], inspired by the score informed source separation approaches [5], we proposed a Masked NMF method, which adapts NMF to the problem of learning meaningful bird sound representations from such noisy, weakly labeled data. Using the weak labels, we constrained parts of the activation matrix to zero, hence obtaining more robust set of bases for sounds of interest and noise.

Masked NMF proved successful for Bird Audio Detection [4], but we observed that often the background sounds were reconstructed using the dictionary of bird sounds. That suggests that constraining the activation matrix is not enough to produce separate set of bases for the sound of interest and the background sounds. Therefore, the difference between the set of bases of the target sound and noise has to be increased by "pushing" the subspaces of the bases apart from each other. In this work, we propose to achieve this by introducing an orthogonality regularization term in the objective function of NMF. Regularization of NMF has been proven

The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n 642685 MacSeNet. MDP is also partly supported by EPSRC grant EP/N014111/1

useful in a number of application. For instance, a temporal constraint improves audio source separation [6], sparsity has been shown to improve NMF performance on real life audio event detection [7] and co-occurrence constraint was used for automatic music transcription [8]. In this paper, inspired by the idea of forcing two dictionaries to be different originally introduced for single channel source separation problem [9], we propose to add an orthogonality regularizer, that decorrelates the two dictionaries and promotes orthogonality between dictionaries of the sound and of the background, resulting in the Orthogonality-Regularized Masked NMF (ORM-NMF) method.

The paper is organized as follows. Sections 2 and 3 introduce the standard NMF and Masked NMF respectively. Then, in Section 4 we describe the proposed method and the derived multiplicative update rules. Later we test the method on a rare event detection task using weakly labeled data described in Section 5. The results are presented in Section 6 followed by the conclusions in Section 7.

2. NON-NEGATIVE MATRIX FACTORIZATION

The goal of NMF is to approximate a non-negative data matrix, typically a time-frequency representation of a given sound, $\mathbf{V} \in {\mathbb{R}_+}^{F \times T}$ as a product of a dictionary $\mathbf{W} \in {\mathbb{R}_+}^{F \times K}$ and its activation matrix $\mathbf{H} \in {\mathbb{R}_+}^{K \times T}$, such that:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}. \tag{1}$$

$\mathbf{W}$ and $\mathbf{H}$ are estimated to minimize some divergence metric $D(\mathbf{V}|\mathbf{WH})$. For any two matrices X and Y, we define $D(\mathbf{X}|\mathbf{Y}) = \sum_{m,n} D(x_{mn}, y_{mn})$. In this work we choose the squared Euclidean distance as the divergence metric, defined as

$$D(\mathbf{V}|\mathbf{WH}) = ||\mathbf{V} - \mathbf{WH}||^2 \tag{2}$$

although other error approximation functions, such as generalized Kullback-Leibler (KL) divergence or Itakura-Saito (IS) divergence [10] are also sensible choices and the proposed method can be easily extended to use those. Euclidean distance can be minimized by alternately updating $\mathbf{W}$ and $\mathbf{H}$ by the following multiplicative update rules [11]:

$$\begin{aligned} \mathbf{W} &\leftarrow \mathbf{W} \odot \frac{\mathbf{VH}^\intercal}{\mathbf{WHH}^\intercal} \\ \mathbf{H} &\leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\intercal\mathbf{V}}{\mathbf{W}^\intercal\mathbf{WH}}, \end{aligned} \tag{3}$$

where $\mathbf{A} \odot \mathbf{B}$ denotes a Hadamard (element-wise) product of two matrices, $\frac{\mathbf{A}}{\mathbf{B}}$ denotes Hadamard division and other multiplications are matrix multiplications.

3. MASKED NMF

In [4] we proposed to extend a standard NMF approach to learning on weakly labeled data. To explain the idea, let us consider the task of detection of rare sound events, as proposed in the Detection and Classification of Acoustic Scenes and Events challenge DCASE2017 [12]. Let $y \in \{0, 1\}$ be a weak label denoting absence or presence of the target sound, $\mathbf{V}^0 = \mathbf{V}^0_1, \cdots, \mathbf{V}^0_{M_0}$ is a set of M_0 training examples with absence of the target sound and $\mathbf{V}^1 = \mathbf{V}^1_1, \cdots, \mathbf{V}^1_{M_1}$ is a set of M_1 training examples with the presence of the target sound. As the data is weakly labeled, examples containing the target sound most probably also contain noise and other sounds. Therefore, we assume that to reconstruct well the target sound training examples ($\mathbf{V}^1$) we also need elements from dictionaries extracted from background sounds examples ($\mathbf{V}^0$). At the same time, we do not expect elements of the dictionary atoms of target sounds to be used for reconstructing $\mathbf{V}^0$. We impose this constraint in the training phase by applying a binary mask to the activation matrix as follows:

$$\begin{aligned} \mathbf{V} = [\mathbf{V}_0, \mathbf{V}_1] &\approx [\mathbf{W}_0, \mathbf{W}_1] \left(\begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \odot \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{01} \\ \mathbf{H}_{10} & \mathbf{H}_{11} \end{bmatrix} \right) \\ &= [\mathbf{W}_0, \mathbf{W}_1] \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{01} \\ \mathbf{0} & \mathbf{H}_{11} \end{bmatrix} \\ &= [\mathbf{W}_0\mathbf{H}_{00}, \mathbf{W}_0\mathbf{H}_{01} + \mathbf{W}_1\mathbf{H}_{11}] \end{aligned} \tag{4}$$

where $\mathbf{W}^0 \in {\mathbb{R}_+}^{F \times K^0}$, $\mathbf{W}^1 \in {\mathbb{R}_+}^{F \times K^1}$ are "sound" and "background" dictionaries respectively, K^0 and K^1 are their corresponding ranks. $\mathbf{0}$ is a matrix of zeros with K^1 rows and the number of columns corresponding to the total size of M_0 background training data, while $\mathbf{1}$ denotes matrices of appropriate dimensions with all elements equal to 1. $\mathbf{H}^{00}$, $\mathbf{H}^{01}$, $\mathbf{H}^{10}$ and $\mathbf{H}^{11}$ are parts of the activation matrix of suitable dimensions. Hence, we are seeking to minimize the Euclidean distance:

$$\begin{aligned} &\min_{\mathbf{W}_0, \mathbf{W}_1, \mathbf{H} \geq 0} \|\mathbf{V} - \mathbf{WH}\|^2 \\ &= \|\mathbf{V}_0 - \mathbf{W}_0\mathbf{H}_{00}\|^2 + \|\mathbf{V}_1 - \mathbf{W}_0\mathbf{H}_{01} - \mathbf{W}_1\mathbf{H}_{11}\|^2 \end{aligned} \tag{5}$$

The masking is implemented through appropriate initialization of the activation matrix. As the update rules of NMF are multiplicative, elements initialized with 0 remain 0 throughout the training. Hence, applying the multiplicative rules from eq. 3 we obtain the dictionary:

$$\mathbf{W} = \left[\mathbf{W}^0, \mathbf{W}^1\right], \tag{6}$$

that was later used for audio classification.

4. PROPOSED METHOD

Masked NMF, although suitable for audio classification task, has some limitations. In previous experiments we have observed that, in spite of the constraint on the activation matrix, the background sounds were often reconstructed using the dictionary of target sounds. It might suggest that the dictionaries are correlated and hence, not discriminative between

the target and background sounds. To overcome this problem we propose to add an additional orthogonality regularizer, which decorrelates the dictionaries by "pushing" them apart from each other. To achieve this, we measure the correlation between the dictionaries $\mathbf{W}_0$ and $\mathbf{W}_1$ using the dot product between them, i.e. $\|\mathbf{W}_1^\intercal \mathbf{W}_0\|^2$, and aim to minimize it. The emphasis, that is given to the orthogonality regularizer can be adjusted by choosing an arbitrary value of λ. Combining the constraint on the activation matrix and the orthogonality regularizer results in the following cost function to minimize:

$$\begin{aligned}&\min_{\mathbf{W}_0,\mathbf{W}_1,\mathbf{H}\geq 0} \|\mathbf{V}-\mathbf{W}\mathbf{H}\|^2 + \lambda\|\mathbf{W}_1^\intercal\mathbf{W}_0\|^2 \\ &= \|\mathbf{V}_0 - \mathbf{W}_0\mathbf{H}_{00}\|^2 + \|\mathbf{V}_1 - \mathbf{W}_0\mathbf{H}_{01} - \mathbf{W}_1\mathbf{H}_{11}\|^2 \\ &+ \lambda\|\mathbf{W}_1^\intercal\mathbf{W}_0\|^2\end{aligned} \tag{7}$$

As $\|\mathbf{W}_1^\intercal\mathbf{W}_0\|^2$ is convex in $\mathbf{W}_0$ and $\mathbf{W}_1$, we can minimize the cost function using the gradient decent. Then, following the derivations of Lee and Sung [11], we obtain the corresponding multiplicative update rules for $\mathbf{W}_0$ and $\mathbf{W}_1$:

$$\begin{aligned}\mathbf{W}_0 &\leftarrow \mathbf{W}_0 \odot \frac{\mathbf{V}_1\mathbf{H}_{01}{}^\intercal + \mathbf{V}_0\mathbf{H}_{00}{}^\intercal}{\mathbf{W}_0\mathbf{H}_{01}\mathbf{H}_{01}{}^\intercal + \mathbf{W}_1\mathbf{H}_{11}\mathbf{H}_{01}{}^\intercal + \lambda\mathbf{W}_1\mathbf{W}_1{}^\intercal\mathbf{W}_0} \\ \mathbf{W}_1 &\leftarrow \mathbf{W}_1 \odot \frac{\mathbf{V}_1\mathbf{H}_{11}{}^\intercal}{\mathbf{W}_0\mathbf{H}_{01}\mathbf{H}_{11}{}^\intercal + \mathbf{W}_1\mathbf{H}_{11}\mathbf{H}_{11}{}^\intercal + \lambda\mathbf{W}_0\mathbf{W}_0{}^\intercal\mathbf{W}_1}\end{aligned} \tag{8}$$

As the regularizer does not influence the activation matrix $\mathbf{H}$, the update rule for $\mathbf{H}$ remains the same as in the original NMF problem formulation shown in Section 2:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\intercal\mathbf{V}}{\mathbf{W}^\intercal\mathbf{W}\mathbf{H}}. \tag{9}$$

5. EXPERIMENTAL SETUP

The proposed method is evaluated on the task of Detection of rare sound events using only weakly labeled data from the audio recordings of the TUT Rare Sound Events 2017. The dataset was provided for Task 2 of the DCASE2017 challenge [12]. All audio files are resampled to sampling rate of 16000 Hz in order to reduce the dimensionality of the data. We extract perceptually motivated mel-spectrograms with 40 components, using a window size of 64 ms, hop size of the same duration. Mel-spectrograms are a common choice for representation of environmental audio [13, 1]. In order to model temporal dynamics of environmental sounds we choose a spectro-temporal representation of the data, which is achieved by grouping several consecutive frames into 2D patches, also known as *shingling*. In our experiments, we set the number of consecutive frames to 4, the value that was chosen empirically.

5.1. Dataset

The dataset consists of around 100 isolated sound examples for three target classes: gunshot, baby crying and glass breaking, together with background audio which is part of the TUT Acoustic Scenes 2016 dataset [14]. For the scenario of audio event detection using weakly labeled data we do not use the isolated recordings for training but we create mixtures with the background audio of equal loudness, i.e. with the 0dB Signal to Noise Ratio (SNR). Each mixture is 4 second long. It is important to reiterate, that we do not know the timestamp of the event in the mix, just a binary label determining weather the mix contains the sound of interest. We evaluate the method on two scenarios:

- *Vanilla scenario* We use 100 test mixtures of 4 second length. The mixtures are created using sound event and background recordings not used in training dataset, mixed with equal loudness (0dB SNR).
- *Challenge scenario* We use 500 mixes of -6dB, 0dB and 6dB SNR of the sounds and backgrounds not used in the training set. The testing mixtures are provided by the organizers of the DCASE2017 challenge. Each testing mixture is 30 second long.

Table 1 shows the number of isolated recordings used to create training and testing mixtures. Equivalent number of audio files not containing the sound was used for training as well as testing in the vanilla scenario.

Table 1. Experimental dataset. Number of sound recordings per class used for training and testing.

Event type	training	testing
Gunshot	134	53
Glass breaking	96	43
Baby crying	106	42

5.2. Event detection

We use the proposed method and the Masked NMF to extract dictionaries $\mathbf{W}_0$ and $\mathbf{W}_1$. In the event detection phase, a test sample is decomposed using the trained dictionaries as follows:

$$\mathbf{V}_{test} = \left[\mathbf{W}^0, \mathbf{W}^1\right] \begin{bmatrix}\mathbf{H}_0 \\ \mathbf{H}_1\end{bmatrix} \tag{10}$$

Finally, $\mathbf{H}_1$ is binarized using a threshold equal to 50% of the maximum value of the entire activation matrix ($\mathbf{H}_0$ and $\mathbf{H}_1$) . The columns of the binarized $\mathbf{H}_1$ that are greater than 0 indicate the presence of the event.

5.3. Evaluation metrics

To evaluate the method we use metrics used in the DCASE2017 challenge, i.e. event-based error rate (ER) and event-based F-score. An event is considered correctly detected using onset-only condition with a collar of 500 ms. The ER is

calculated by adding the number of substitutions, insertions and deletions for each class before dividing it by the total number of events. The F-score is computed as the harmonic mean between precision and recall based on the total amount of false negatives, true positives and false positives per class. We refer the reader to [15] for more details and explanations about these metrics.

6. RESULTS

The parameters of the system, namely K_0, K_1 and λ were chosen empirically during the development of the method and set constant throughout the experiments to allow for a meaningful comparison. The parameters were chosen to be: $K_0 = 50$, $K_1 = 10$ and $\lambda = 1000$.

6.1. Vanilla scenario

We compare the proposed method with the Masked NMF approach. We can see that for 4 seconds chunks of audio with 0 dB SNR

Table 2. Evaluation results in Vanilla Scenario. Error Rate (ER) and F-score (F1) are reported for the proposed method and Masked NMF.

Event type	**Proposed**		Masked NMF	
	ER	F1	ER	F1
Gunshot	0.26	87.5%	0.30	86.4%
Glass breaking	0.21	89.4%	0.18	90.9%
Baby crying	0.85	51.3%	0.92	51.8%

6.2. Challenge Scenario

To show that the method has a potential to be used in more complicated scenarios, we evaluate the trained models on the official development data of the DCASE2017 Challenge. Table 3 shows the results for the proposed method, Masked NMF and the DCASE2017 baseline.

Table 3. Evaluation results in Challenge scenario. Error Rate (ER) and F-score (F1) are reported for the proposed method, Masked NMF and DCASE2017 baseline.

Event type	**Proposed**		Masked NMF		DCASE	
	ER	F1	ER	F1	ER	F1
Gunshot	0.79	64.7%	0.81	64.2%	0.69	57.4%
Glass breaking	0.87	50.1%	0.94	52.5%	0.22	88.5%
Baby crying	0.97	39.1%	1.07	37.9%	0.67	72.0%

6.3. Discussion

The results in Table 2 show that the proposed method is a promising way to learn on weakly labeled data. It is interesting to see that the performance on *gunshot* and *glass breaking* sounds is much higher than on *baby crying* sounds. This may show that the proposed method is more suitable for detection of impact than harmonic sounds. Moreover, separate parameter tuning for each class could be beneficial. Further analysis is needed to investigate the reasons for such a big difference.

The results in Table 3 confirm our findings using the Vanilla scenario. The method performs well on the *gunshot* detection, reasonably well on *glass breaking* detection and much worse on *baby crying* detection. It has to be reiterated, that the baseline for DCASE 2017 was using strongly annotated data, hence we expected our method to perform worse, as we allowed ourselves to use weakly labelled data only.

From both scenarios we can see that regularization lowers the Error Rate, but F-score is not always increased. The reasons for this behaviour need more investigation. However, it can be concluded by analysing the results of the DCASE 2017 Challenge, that not always methods that achieve the lower ER achieve higher F-score.

7. CONCLUSIONS

We proposed a novel method based on NMF for learning sound representations on weakly labeled data. Adding a constraint on the activation matrix and an orthogonality regularizer to the standard NMF formulation we are able to learn sound representations without isolated or strongly annotated training data. Using the task of detection of rare events as an example, we showed that the method is a promising direction for Audio Event Detection when no isolated or annotated sounds are present. However, the performance of the method strongly depends on the type of target sound.

In future we plan to compare our method with other algorithms tailored especially for weakly labeled data. Moreover, we would like to understand better the influence of the parameters of the system and the reasons for a big discrepancy of the results between different classes. Finally, we want to compare iterative methods for decorrelating the dictionaries with the proposed approach.

8. REFERENCES

[1] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*, 2017, pp. 6445–6449.

[2] Q. Zhou and Z. Feng, "Robust sound event detection through noise estimation and source separation using

NMF," Tech. Rep., DCASE2017 Challenge, September 2017.

[3] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2013)*, 2013, pp. 5–8.

[4] I. Sobieraj, Q. Kong, and M. D. Plumbley, "Masked non-negative matrix factorization for bird detection using weakly labeled data," in *Proceedings of the 25th European Signal Processing Conference (EUSIPCO 2017)*, 2017.

[5] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012, pp. 129–132.

[6] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proceedings of the International Computer Music Conference (ICMC 2003)*, 2003, pp. 231–234.

[7] I. Sobieraj and M. D. Plumbley, "Coupled sparse NMF vs. random forest classification for real life acoustic event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 90–94.

[8] S. K. Tjoa and K. J. R. Liu, "Multiplicative update rules for nonnegative matrix factorization with co-occurrence constraints," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, Mar. 2010, pp. 449–452.

[9] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation.," in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH2013)*, 2013.

[10] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, September 2011.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[12] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[13] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," Tech. Rep., DCASE2017 Challenge, September 2017.

[14] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proceedings of the 24rd European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.