Masked Non-negative Matrix Factorization for Bird Detection Using Weakly Labeled Data

Iwona Sobieraj, Qiuqiang Kong, Mark D. Plumbley
University of Surrey
Centre for Vision Speech and Signal Processing
Guildford, Surrey GU2 7XH, United Kingdom
{i.sobieraj, q.kong, m.plumbley}@surrey.ac.uk

***Abstract*—Acoustic monitoring of bird species is an increasingly important field in signal processing. Many available bird sound datasets do not contain exact timestamp of the bird call but have a coarse weak label instead. Traditional Non-negative Matrix Factorization (NMF) models are not well designed to deal with weakly labeled data. In this paper we propose a novel Masked Non-negative Matrix Factorization (Masked NMF) approach for bird detection using weakly labeled data. During dictionary extraction we introduce a binary mask on the activation matrix. In that way we are able to control which parts of dictionary are used to reconstruct the training data. We compare our method with conventional NMF approaches and current state of the art methods. The proposed method outperforms the NMF baseline and offers a parsimonious model for bird detection on weakly labeled data. Moreover, to our knowledge, the proposed Masked NMF achieved the best result among non-deep learning methods on a test dataset used for the recent Bird Audio Detection Challenge.**

I. Introduction

Automatic detection of bird sounds is an important task for many applications, such as environmental monitoring [1] or audio indexing. With appearance of publicly available audio datasets, such as freefield1010 [2], Warblr [3] or Chernobyl dataset from TREE project[1] and challenges for bird recognition, such as the LifeCLEF Bird Identification Task [4] and the recent Bird Audio Detection (BAD) Challenge [3], the problem has received considerable attention from audio research community.

The complete goal of bird detection is to be able to detect the beginning and end of a bird call and classify it from an arbitrary number of species. However, a simple task of determining the presence or absence of a bird sound in an audio file is often a first step in analysis of large audio datasets. Although determining the presence of a bird is a straightforward binary classification task, the classifier has to cope with several difficulties. Firstly, the models need to be able to recognize an unspecified variety of bird calls. Secondly, as data annotation is an expensive and time-consuming task, available data is often *weakly labeled*. This means that we are provided with coarse labels determining the presence or absence of a bird in an audio recording, rather than an exact timestamp of the bird call or even the number of times it has occurred [5]. Finally, real-life recordings contain high amount of noise, hence robust methods are necessary.

Many state of the art methods for bird detection are based on Deep Learning techniques [6]. Although these provide excellent results, deep learning models are complex and require a large number of parameters to optimize. Moreover, they rely on huge amount of training data, which is often hard to acquire in audio domain. In contrast, Non-negative Matrix Factorization (NMF) methods offer parsimonious models and are known to provide meaningful and interpretable decompositions of audio data into parts [7]. NMF can be interpreted as a feature learning method, which aim is to learn a transformation that, when applied to data, increases the performance of a given system. Feature learning has been shown to be beneficial for large-scale bird sounds classification [2]. At the same time, NMF has been successfully used in a number of acoustic event detection tasks [8], [9], [10]. Typically, NMF is used in an unsupervised manner to extract separate dictionaries for each class from its isolated recordings. The activations of the dictionary elements determine then the presence of events [11], [9]. However, in real-life applications we often do not have access to isolated sounds or exact labels determining the timestamp of each event, but to weakly labeled data only, for which NMF seems to be ill-suited.

In this paper we propose a *Masked NMF* approach for extracting dictionaries from weakly labeled data. The method is inspired by score-informed musical source separation methods, where a MIDI musical score is used as a constraint for the musical note activation matrix [12]. A similar approach has been proposed for polyphonic acoustic event detection, where activity annotations served as a constraint [13]. However, in both cases the data is well annotated, hence the exact appearance of each note/sound event is known. We adapt the method for weakly labeled data using the coarse labels as a constraint on the activation matrix of NMF decomposition. More specifically, we jointly train dictionaries for "bird" and "non-bird" sounds by applying a binary mask on the activation matrix of NMF in the training phase. In that way, we allow the reconstruction of bird sounds using both bird and non-bird dictionaries but we set to zero activations of bird dictionary atoms during the reconstruction of non-bird sounds. Hence,

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/NO14111/1

[1]http://tree.ceh.ac.uk/

non-bird sounds are reconstructed with non-bird dictionary only. The proposed *Masked NMF* method is evaluated and compared to the state of the art methods in the BAD Challenge setup.

II. Background

A. Non-negative Matrix Factorization (NMF)

NMF is a well known technique to decompose non-negative data into a product of two non-negative low rank matrices, which has shown to be beneficial in audio processing field. The goal of NMF is to approximate a 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 an arbitrary cost function $D(\mathbf{V}|\mathbf{WH})$. The popular choice of a cost function is the generalized Kullback-Leibler (KL) divergence,

$$D(x|y) = x\log\frac{x}{y} - x + y \tag{2}$$

although other error approximation functions, such as Euclidean distance or Itakura-Saito (IS) divergence are also a sensible choice [14]. KL divergence is minimized by alternately updating $\mathbf{W}$ and $\mathbf{H}$ by the following multiplicative update rules[7]:

$$\begin{aligned} \mathbf{W} &\leftarrow \mathbf{W} \odot \frac{\mathbf{V}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T} \\ \mathbf{H} &\leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T\mathbf{1}}, \end{aligned} \tag{3}$$

where 1 is a matrix of dimensions $F \times T$ with all its elements equal to 1, $\mathbf{A} \odot \mathbf{B}$ denotes a Hadamard (element-wise) product of two matrices, $\frac{\mathbf{A}}{\mathbf{B}}$ - Hadamard division and other multiplications are matrix multiplications.

B. Non-negative Matrix Factorization for audio detection

1) Unsupervised NMF (UNMF): The straightforward application of NMF for binary audio classification is to use it as an unsupervised feature learning method. The approach is similar to the one proposed by Stowell and Plumbley [15] but instead of learning an over-complete dictionary we extract a parsimonious representation using NMF. In that case we concatenate all the training examples from both bird and non-bird examples into one matrix $\mathbf{V}$ and extract a single dictionary.

2) Class-conditioned NMF (CCNMF): Alternatively, NMF is used to extract separate dictionaries for each of N classes [11], [16], [9]. M training examples for each class are concatenated and jointly factorized. For a binary classification task, $N = 2$ and the training is performed as follows:

$$\begin{aligned} \left[\mathbf{V}_1^0, \mathbf{V}_2^0, \cdots, \mathbf{V}_{M_0}^0\right] &\approx \mathbf{W}^0\mathbf{H}^0 \\ \left[\mathbf{V}_1^1, \mathbf{V}_2^1, \cdots, \mathbf{V}_{M_1}^1\right] &\approx \mathbf{W}^1\mathbf{H}^1. \end{aligned} \tag{4}$$

In the detection phase, all dictionaries are typically concatenated and the activation matrix is retrieved:

$$\mathbf{V}_{\text{test}} \approx \left[\mathbf{W}^0, \mathbf{W}^1\right] \begin{bmatrix} \mathbf{H}^0 \\ \mathbf{H}^1 \end{bmatrix} = \mathbf{W}^0\mathbf{H}^0 + \mathbf{W}^1\mathbf{H}^1. \tag{5}$$

Usually, the training data is formed of isolated recordings, therefore the N extracted dictionaries are expected to be discriminative for each of class. However, it has been shown that the dictionaries learnt from different events contain similar elements [9]. That means that a given audio event is reconstructed not only using elements from its corresponding dictionary but also atoms from dictionaries of other events, which makes discriminative classification more difficult. This problem may appear even more clearly, when the training data contains noise common for several classes, which is the case for weakly labeled data.

3) Event detection: In order to determine which event was present, thresholding may be applied to the activation matrix [10], [13]. Alternatively, if we do not need to locate the sound in time, the activation matrix can used as an input to an arbitrary classifier, such as SVM or random forest [13], [15].

III. Proposed method

To alleviate the limitation of the standard NMF not leading to discriminative dictionaries, we propose to add a constraint on the activation matrix in the dictionary generation phase. In this section we describe in detail the proposed method.

A. Dictionary Generation

Let us consider a binary classification case of bird detection, where the number of classes is $N = 2$, and where $y \in \{0, 1\}$ is a label denoting absence or presence of a bird, $\mathbf{V}^0 = \mathbf{V}_1^0, \cdots, \mathbf{V}_{M_0}^0$ is a set of M_0 training examples with absence of birds and $\mathbf{V}^1 = \mathbf{V}_1^1, \cdots, \mathbf{V}_{M_1}^1$ is a set of M_1 training examples with the presence of birds. As the data is weakly labeled, examples containing bird song most probably also contain noise and other sounds. Therefore, we assume that to reconstruct well bird training examples ($\mathbf{V}^1$) we also need elements from dictionaries extracted from non-bird examples ($\mathbf{V}^0$). At the same time, we do not expect elements of the dictionary atoms of bird sounds to be present 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:

$$\begin{aligned} \left[\mathbf{V}^0, \mathbf{V}^1\right] \approx \\ \approx \left[\mathbf{W}^0, \mathbf{W}^1\right] \left(\begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \odot \begin{bmatrix} \mathbf{H}^0 \\ \mathbf{H}^1 \end{bmatrix} \right) = \\ = \left[\mathbf{W}^0, \mathbf{W}^1\right] \begin{bmatrix} \mathbf{H}^{00} & \mathbf{H}^{01} \\ \mathbf{0} & \mathbf{H}^{11} \end{bmatrix} = \\ = \left[\mathbf{W}^0\mathbf{H}^{00}, \mathbf{W}^0\mathbf{H}^{01} + \mathbf{W}^1\mathbf{H}^{11}\right] \end{aligned} \tag{6}$$

where $\mathbf{W}^0 \in {\mathbb{R}_+}^{F \times K^0}$, $\mathbf{W}^1 \in {\mathbb{R}_+}^{F \times K^1}$ are "bird" and "non-bird" 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

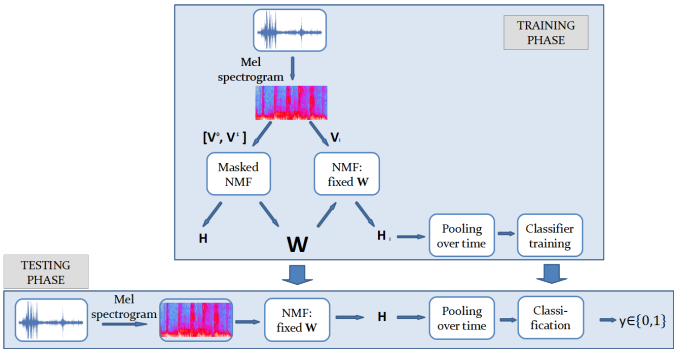

Fig. 1. A simplified block diagram of the bird audio detection system. See text for the complete description of the system.

of M_0 non-bird training data, while $\mathbf{1}$ denotes matrices of appropriate dimensions with all elements equal to 1. $\mathbf{H}^{00}$, $\mathbf{H}^{01}$ and $\mathbf{H}^{11}$ are parts of the activation matrix of suitable dimensions. 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. In this way we obtain a dictionary:

$$\mathbf{W} = \left[\mathbf{W}^0, \mathbf{W}^1\right] \tag{7}$$

which we then use for bird detection.

B. Bird Detection

Having extracted the dictionary $\mathbf{W}$ by the masked NMF (eq. (6)) we fix $\mathbf{W}$ and use to decompose each of the training examples via standard NMF (eq. (1)). We obtain an activation matrix $\mathbf{H}$ for each file. As the data contains recordings of different lengths, and to reduce the dimensionality of the representations, we summarize (pool) the activation matrices over the time. We choose the mean and standard deviation of each row of $\mathbf{H}$, as this combination of pooling functions has shown beneficial for audio classification [17], [15]. We then use the pooled activation matrices as feature vectors to train a binary classifier. In this paper we choose a random forest classifier [18] with 500 trees, which has been previously successfully used for audio classification [17], [15].

For a newly observed signal $\mathbf{V}_{\text{test}}$ we follow the same procedure: we decompose the signal using $\mathbf{W}$, pool the obtained activation matrix $\mathbf{H}_{\text{test}}$ and classify with the trained binary classifier.

A simplified block diagram of the bird detection system using the proposed Masked NMF is shown in Figure 1.

IV. Experiments

A. Datasets and metrics

For the experiments we use three datasets provided by the organisers of the BAD Challenge: Warblr [3], freefield1010 [2] and the Chernobyl dataset from TREE project[2]. Warblr is a crowd-sourced dataset that consists of 10,000 ten-second real-life audio signals which contain different background noise and also fake bird calls imitated by humans. Freefield1010 is a collection of over 7,000 excerpts from field recordings around the world, gathered by the FreeSound project[3]. The third dataset, Chernobyl, is a part of approximately 10,000 hours of audio captured in Chernobyl Exclusion Zone (CEZ) [3]. This dataset is kept unpublished by the organizers of the Bird Detection Challenge for the sake of future evaluations on unknown data. We perform experiments in two scenarios:

1) Development scenario: we use 10% of Warblr and 10% of freefield1010 datasets for training, and another 10% of each of the two datasets for testing. We decided to use just a subset of all available data to allow for faster development.
2) Deployment scenario: we use entire Warbl and freefield1010 datasets for training with the parameters chosen during development, we test the model on test data provided by the organisers of the BAD Challenge via an online submission system[4]. The test data consist of 1293 files from freefield1010 and Chernobyl datasets.

We evaluate the results using widely used evaluation metric: Receiver Operator Characteristics curves and a corresponding Area Under the ROC Curve (AUC) [19].

B. Data representation

As a data representation we use normalized mel spectrograms extracted with 40 bands. We use a Hamming window of 23 ms with no overlap on signals sampled with 44.1 kHz. Then, we group several (N_{sh}) consecutive frames by concatenating them into a single larger vector. This is known as *shingling* and allows for learning the temporal information, which has been shown to be beneficial for environmental audio classification [17].

C. Masked NMF performance

We compare the proposed Masked NMF with the two baseline methods: CCNMF and UNMF (see Section II-B). For a meaningful comparison we set the total size of the dictionary to be equal for each of the methods. i.e. $K = 60$. For Masked NMF and CCNMF we use $K^1 = 10$ bases for bird sounds and $K^0 = 50$ for non-bird sounds. We fix $N_{sh} = 4$. The number of bases was chosen after a search through a total range from 20 to 100 bases.

Figure 2 shows the ROC curves for the Masked NMF and both baseline methods in the development experimental scenario. Table I shows the corresponding AUC scores. The proposed method outperforms both baseline approaches.

D. Length of temporal context

We also investigated the influence of the number of consecutive frames (shingles) taken as an input vector. Figure 3 shows the ROC curves for different numbers of shingles. From the set $N_{sh} \in 1, 4, 12$ we found $N_{sh} = 4$ to give the best performance.

[2]http://tree.ceh.ac.uk/

[3]https://freesound.org/

[4]http://lsis-argo.lsis.org/

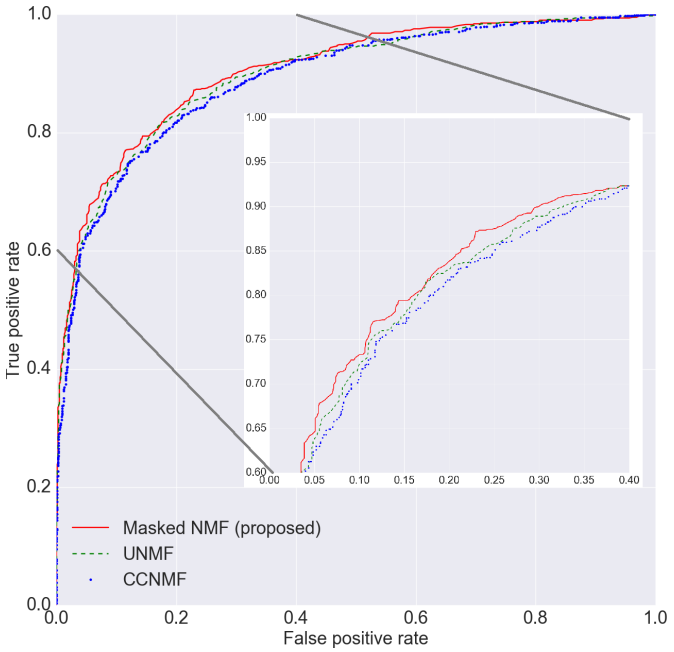

Fig. 2. Receiver Operator Characteristics (ROC) curves for different NMF methods. We compare Masked NMF, Class-conditioned NMF (CCNMF) and unsupervised NMF (UNMF) of the same dictionary size, i.e with 60 bases (10 for bird sounds, 50 for non-bird sounds). The inset shows a magnified left upper corner of the ROC curve.

Method	AUC
UNMF	89.8%
CCNMF	89.1%
Masked NMF (Proposed)	**90.4%**

TABLE I
AREA UNDER THE CURVE (AUC) SCORE FOR DIFFERENT NMF METHODS: MASKED NMF (PROPOSED), CLASS-CONDITIONED NMF (CCNMF) AND UNSUPERVISED NMF (UNMF) OF THE SAME DICTIONARY SIZE, I.E WITH 60 BASES (10 FOR BIRD SOUNDS, 50 FOR NON-BIRD SOUNDS)

Number of consecutive frames	AUC
1 frame	89.4%
4 frames	**90.4%**
12 frames	89.7%

TABLE II
AREA UNDER THE CURVE (AUC) SCORE FOR DIFFERENT NUMBER OF CONSECUTIVE FRAMES USED AS DATA REPRESENTATION: 1, 4 AND 12 FRAMES USING MASKED NMF WITH 60 BASES (10 FOR BIRD SOUNDS, 50 FOR NON-BIRD SOUNDS)

E. Bird Audio Detection Challenge result

In order to compare Masked NMF with other state of the art methods we trained the models using all available data from BAD Challenge setup (see IV-A: Deployment Scenario). We chose our best performing model fixing the parameters to the following: $K^0 = 50$, $K^1 = 10$, $N_{sh} = 4$. We submitted the results on the test dataset to the online submission system[5], which reports the AUC score for the test dataset. The system evaluates the results on a preview of 1293 audio files from

[5]http://lsis-argo.lsis.org/

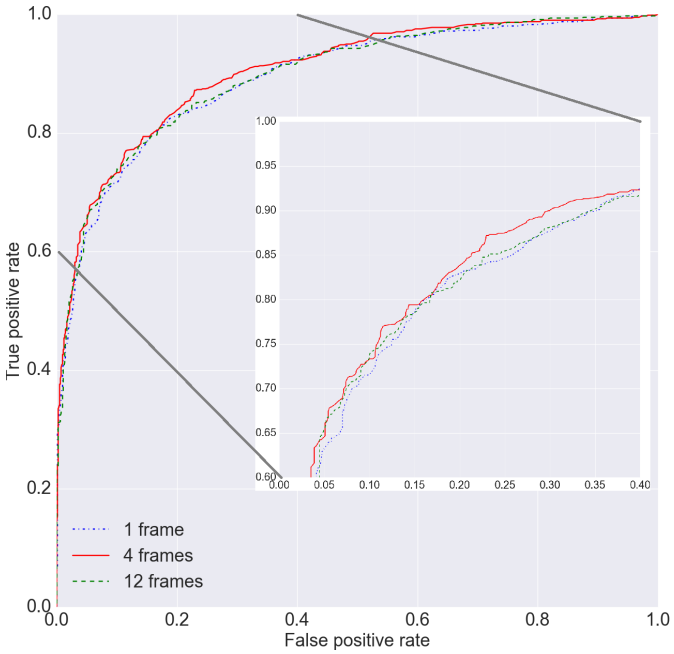

Fig. 3. Receiver Operator Characteristics (ROC) curves for different number of concatenated frames (shingle size). We use the proposed Masked NMF with 60 bases (10 for bird sounds, 50 for non-bird sounds). The inset shows a magnified left upper corner of the ROC curve.

Method	AUC
CNN [20]	88.9%
CRNN [21]	88.3%
Masked NMF (Proposed)	**80.1%**
CCNMF (baseline)	78.3%
UNMF (baseline)	77.8%
GMM/SVM [22]	77.2%

TABLE III
AREA UNDER THE CURVE (AUC) SCORE REPORTED BY THE OFFICIAL SUBMISSION SYSTEM OF THE BIRD DETECTION CHALLENGE FOR SIX SYSTEMS: PROPOSED MASKED NMF, TWO BEST PERFORMING DEEP LEARNING SYSTEMS (CNN AND CRNN), THE BEST NON-DEEP LEARNING SYSTEM (GMM/SVM) AND TWO BASELINE METHODS (SEE II-B)

the dataset, which, according to the organizers of the BAD Challenge, gives a sound estimate of the performance on the whole testing dataset. To be consistent, we report the results of the other methods on the preview dataset as well [6].

Table III shows the final performance reported by the submission system of BAD Challenge. We compare the method with two best deep learning approaches (CNN and CRNN) and the best non-deep learning method (GMM/SVM) [6]. Masked NMF achieves the best performance among non-deep learning methods, although 8 percentage points lower than the best deep learning approach.

V. CONCLUSIONS

This paper proposes a new method for bird audio detection using weakly labeled data, Masked NMF. The proposed method incorporates information from weak labels by adding a constraint on the activation matrix during the training step of NMF. A binary mask is applied on the activation matrix

to allow reconstruction of bird sounds using both"bird" and "non-bird" dictionaries and reconstruction of non-bird sounds using non-bird dictionaries only. Dictionaries extracted using the Masked NMF method achieve better performance on bird audio detection task than dictionaries extracted using unsupervised or class-conditioned NMF. The proposed method achieved the best result among non Deep Learning methods on the Bird Audio Detection Challenge test data.

The method has the potential to be extended to different types of sounds and multiple classes which we will do in the nearest future. Moreover, inducing different levels of sparsity on the activations of NMF for each of the dictionaries is an interesting approach that we will investigate. In the next work we will further investigate the influence of the rank of decomposition on the extracted dictionaries.

REFERENCES

[1] R. T. Buxton and I. L. Jones, "Measuring nocturnal seabird activity and status using acoustic recording devices: Applications for island restoration," *Journal of Field Ornithology*, vol. 83, no. 1, pp. 47–60, 2012.

[2] D. Stowell and M. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," in *Proceedings of Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, January 2014.

[3] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *Proceedings of the IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[4] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, A. Rauber, and A. Joly, "LifeCLEF Bird Identification Task 2014," in *CLEF: Conference and Labs of the Evaluation Forum*, ser. Information Access Evaluation meets Multilinguality, Multimodality, and Interaction, Sheffield, United Kingdom, September 2014.

[5] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*, ser. MM '16. New York, NY, USA: ACM, 2016, pp. 1038–1047.

[6] "Bird audio detection challenge results," Queen Mary University of London, http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge-results/, 2017, retrieved on: 1st of March 2017.

[7] 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.

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

[9] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 45–49.

[10] 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.

[11] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, H. Van, K. Arenberg, T. M. Kempen, and K. U. Leuven, "An exemplar-based NMF approach to audio event detection," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (D-CASE 2013)*, 2013, extended abstract. [Online]. Available: http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/OL/GVV.pdf

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

[13] S. Wang and J. Ortiz, "Non-Negative Matrix Factorization of Signals With Overlapping Events for Event Detection Applications," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*, submitted for publication.

[14] 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.

[15] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, 2014.

[16] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

[17] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 2015, pp. 171–175.

[18] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[20] T. Grill, "Bird audio detection challenge," http://jobim.ofai.at/gitlab/gr/bird_audio_detection_challenge_2017, 2017, retrieved on: 1st of March 2017.

[21] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," Bird Audio Detection Challenge, Tech. Rep., 2017. [Online]. Available: http://machine-listening.eecs.qmul.ac.uk/wp-content/uploads/sites/26/2017/01/cakir.pdf

[22] A. Thakur, J. Jain, P. Rajan, and A. Dileep, "Bird audio detection using probability sequence kernels," Bird Audio Detection Challenge, Tech. Rep., 2017. [Online]. Available: http://machine-listening.eecs.qmul.ac.uk/wp-content/uploads/sites/26/2017/02/badChallenge_iitMandi.pdf