OVERLAPPING SOUND EVENT DETECTION WITH SUPERVISED NONNEGATIVE MATRIX FACTORIZATION

Victor Bisot, Slim Essid, Gaël Richard

LTCI, 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 β-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 β-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

This work was partly funded by the European Union under the FP7-LASIE project (grant 607480)

described in Section 3. Experimental results are presented in Section 4. Finally, conclusions and directions for future work are exposed in Section 5.

2. ACOUSTIC EVENT DETECTION SYSTEM

2.1. Input representation

The NMF stage takes as input the time frequency representation $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ of the audio signals, where F is the number of frequency bands and N is the number of time frames. The nonnegative matrix $\mathbf{V}$ 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 $\mathbf{V}$ such as:

$$\mathbf{V} \approx \mathbf{WH}, \tag{1}$$

with $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$. 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_β, NMF is obtained by solving the following optimization problem:

$$\min_{\mathbf{W},\mathbf{H}} D_\beta(\mathbf{V}|\mathbf{WH}) \text{ s.t. } \mathbf{W}, \mathbf{H} \geq 0. \tag{2}$$

In this work D_β represents the β-divergence [17]. The particular cases of interest for the β-divergence are the Euclidean distance ($\beta = 2$), Kullback-Leibler ($\beta = 1$) and Itakura-Saito ($\beta = 0$). One property of the β 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 β-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 β-divergence reconstruction cost.

The TD-NMF model first considers the optimal projections $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ of the data point $\mathbf{v}$ on the dictionary $\mathbf{W}$. The projections are defined as solutions of the nonnegative projection with ℓ_1 and ℓ_2-norm penalties expressed as in an elastic-net fashion:

$$\mathbf{h}^\star(\mathbf{v}, \mathbf{W}) = \min_{\mathbf{h} \in \mathbb{R}_+^K} D_\beta(\mathbf{V}|\mathbf{Wh}) + \lambda_1 \|\mathbf{h}\|_1 + \lambda_2 \|\mathbf{h}\|_2^2; \tag{3}$$

where λ_1 and λ_2 are nonnegative regularization parameters. Let each data frame $\mathbf{v}$ 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 $l_s(y, \mathbf{A}, \mathbf{h}^\star(\mathbf{v}, \mathbf{W}))$, a function of the optimal projection $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$, where $\mathbf{A} \in \mathcal{A}$ are the parameters of the classifier. The TD-NMF problem is now expressed as a joint minimization of the expected classification cost over $\mathbf{W}$ and $\mathbf{A}$:

$$\min_{\mathbf{W} \in \mathcal{W}, \mathbf{A} \in \mathcal{A}} f(\mathbf{W}, \mathbf{A}) + \frac{\nu}{2} \|\mathbf{A}\|_2^2, \tag{4}$$

with

$$f(\mathbf{W}, \mathbf{A}) = \mathbb{E}_{y,\mathbf{v}}[l_s(y, \mathbf{A}, \mathbf{h}^\star(\mathbf{v}, \mathbf{W}))]. \tag{5}$$

Here, $\mathcal{W}$ is defined as the set of nonnegative dictionaries containing unit l_2-norm basis vectors and ν 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 β case, the expression of the gradient of $f(\mathbf{W}, \mathbf{A})$ with respect to the dictionary $\mathbf{W}$ 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.

Home		**Residential Area**	
Event label	Instances	Event label	Instances
(object) rustling	41	(object) banging	15
(object) snapping	42	bird singing	162
cupboard	27	car passing by	74
cutlery	56	children shouting	23
dishes	94	people speaking	41
drawer	23	people walking	32
glass jingling	26	wind blowing	22
object impact	155		
people walking	24		
washing dishes	60		
water tap running	37		

Table 1: Event labels and number of instances for the two environments of the DCASE 2016 development set.

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 ℓ_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 ν regularization parameter was fixed to $\nu = 10$ after testing.

The TD-NMF model is initialized with the $\mathbf{W}$ and $\mathbf{A}$ 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 β, 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 ℓ_1 and ℓ_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.

			Res. A.		**Home**		**Mean**	
Method	K	β	ER	F1	ER	F1	ER	F1
NMF	8	2	61	54	88	27	74.5	40.5
NMF	8	1	60	58	88	27	74	42.5
NMF	16	2	60	56	87	29	73.5	42.5
NMF	16	1	59	58	86	30	72.5	44
TD-NMF	8	2	56	62	**83**	**37**	**69.5**	49.5
TD-NMF	8	1	58	60	85	34	71.5	47
TD-NMF	16	2	56	**64**	85	36	70.5	**50**
TD-NMF	16	1	**55**	**64**	86	34	70.5	49

Table 2: Error rate and F1 score on 1 second segments for NMF and TD-NMF as a function of the dictionary size K and the divergence β. Results are given for both environments of the DCASE dataset as well as their average. The best results are highlighted in bold text.

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 β parameter, decomposing the data with the Kullback-Leibler divergence ($\beta = 1$) slightly increases the performance over using the Euclidean distance for the unsupervised

	Method	Features	Average. ER	F1
[2]	GMM	MFCC	91	23.7
[25]	Random Forests	MFCC	76	38.5
[12]	GRNN	Spectrogram	73	47.6
[26]	RNN	Mel energy	81.5	**49.8**
	NMF	Mel spectrum	72.5	44
	TD-NMF	Mel spectrum	**69.5**	**49.5**

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.

			Residential Area			
			All segments		with overlap	
Method	K	β	ER	F1	ER	F1
TD-NMF	8	2	61	56	75	37
TD-NMF	8	1	59	58	71	42
TD-NMF	16	2	**58**	**59**	75	38
TD-NMF	16	1	60	57	**70**	**43**

Table 4: Error rate and F1 score on 100 ms segments for TD-NMF in function of the dictionary size K and the divergence β. 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.

NMF. However the $\beta = 1$ case does not present any increase in performance for the TD-NMF, where the discriminative setting appears to learn better dictionaries rather independently of the reconstruction cost. Regarding the different environments, the performance for the *home* indoors recordings is significantly worse than for *residential area*. This can be a consequence of the higher number of labels in the *home* environment as well as some of them corresponding to more abstract concepts like *(object) rustling*. Moreover the system can easily confuse certain labels as some of them overlap by definition. For example *washing dishes* describes an action that is likely to produce *dishes* and *water tap running* events.

In Table 3 we confront our methods to the best submissions on the development set of the DCASE challenge. The first method is the challenge baseline [2], it is a MFCC-GMM approach where a different two class GMM is trained for every label. We also include the 3 best submissions, with two of them being DNN-based systems. The first one uses Gated recurrent neural Networks (GRNN) [12], the second one is based on bi-directional recurrent neural Networks (Bi-RNN) [26] and the last one on Random forests with MFCC features [25]. The average results over the two environments show that both our NMF and TD-NMF systems achieve better results on both metrics compared to the MFCC-GMM baseline. Moreover, even the unsupervised NMF can reach a similar ER as the best system while still having a lower F1 score. With the TD-NMF, our system attains lower ER than all other submissions while keeping a similar F1 score. It is important to note that no definitive conclusions can be drawn for these results as the ranking of the challenge submissions changed a lot when applied on the evaluation set. However they show the potential of the proposed supervised NMF system compared to more complex RNN-based systems. Indeed, the useful spectral information for the event detection can be factorized in a 40×8 dictionary and still achieve competitive performances.

4.5. Performance on overlapping events

In this section we propose to discuss the interest of our system for the specificity of polyphonic AED by changing the context in which we apply the evaluation metrics. The goal is to highlight the capability of an AED system to detect and classify overlapping events. To do so, we propose to keep the same metrics and experimental setting as described in Section 4.3, but instead, we only score on segments that contain more than one label. In practice this is achieved by discarding, for the prediction and annotation, all frames associated with less than two labels in the annotation. Moreover, we compute the ER and F1 scores over shorter segment of 100 ms. This second change to the metric is motivated by the fact that events may overlap over short periods of time. Therefore, scoring on 1 seconds can be very forgiving in this context, making the results less likely to translate the ability of the system to deal with polyphony. Scoring on 100 ms segments also has the advantage of providing a better idea of the temporal precision in the systems predictions. In order to differentiate the effect of the two mentioned changes to the scoring strategy, we also include the results computed over the full set of 100ms segments (including non-overlapping events). This process is particularly relevant for the *residential area* environment as almost 15% of the frames contain overlapping events in the annotation. Whereas it is only the case for 5% of them in the *home* environment.

The results for TD-NMF obtained on the full set of 100 ms segments and on the ones containing overlap are reported in Table 4. First, as expected, the ER and F1 scores on all segments indicate that changing the scoring to 100 ms segments degrades performance in all cases. However, the choice of β does not seem to have a significant effect on the temporal precision of the predictions. In the case of overlapping events, the Kullback-Leibler TD-NMF divergence outperforms the Euclidean case by 4 to 5 points in each metric. This follows the intuition that changing the divergence can be beneficial in the presence of overlap, where the frequency structure of certain events can be masked by others. While the $\beta = 1$ TD-NMF did offer improvements on the full set of segments, this last result indicates that its use should still be considered when dealing with highly polyphonic data.

5. CONCLUSION

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 β-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.

6. REFERENCES

[1] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *Proc. of European Signal Processing Conference*, 2013.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Proc. of European Signal Processing Conference*, 2016.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.

[4] Ya-Ti Peng, Ching-Yung Lin, Ming-Ting Sun, and Kun-Cheng Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *Proc. of International Conference on Multimedia and Expo*. IEEE, 2009, pp. 1218–1221.

[5] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search.," in *Proc. of Interspeech*, 2009, pp. 1151–1154.

[6] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[7] A. Plinge, R. Grzeszick, and G. Fink, "A bag-of-features approach to acoustic event detection," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3704–3708.

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

[9] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 151–155.

[10] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[11] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, 2016, pp. 6–10.

[12] M. Zhrer and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, 2016, pp. 115–119.

[13] A. Gorin, N. Makhazhanov, and N. Shmyrev, "DCASE 2016 sound event detection system based on convolutional neural network," Tech. Rep., DCASE2016 Challenge, 2016.

[14] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[15] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-euclidean sparse nmf via bilevel optimization with applications to speech enhancement," in *Proc. of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2014, pp. 11–15.

[16] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

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

[18] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification," HAL: working paper or preprint (hal-01362864), Sept. 2016.

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

[20] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software.," in *Proc. of ISMIR*, 2010, pp. 441–446.

[21] J. Eggert and E. Körner, "Sparse coding and nmf," in *Proc. of 2004 IEEE International Joint Conference on Neural Networks*, 2004, vol. 4, pp. 2529–2533.

[22] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse nmf–half-baked or well done?," Tech. Rep., Mitsubishi Electric Research Labs (MERL), 2015.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[25] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, 2016, pp. 20–24.

[26] T. H. Vu and J.-C. Wang, "Acoustic scene and event recognition using recurrent neural networks," Tech. Rep., DCASE2016 Challenge, 2016.