COUPLED SPARSE NMF VS. RANDOM FOREST CLASSIFICATION FOR REAL LIFE ACOUSTIC EVENT DETECTION

Iwona Sobieraj *Mark D. Plumbley*

i.sobieraj@surrey.ac.uk m.plumbley@surrey.ac.uk

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

ABSTRACT

In this paper, we propose two methods for polyphonic Acoustic Event Detection (AED) in real life environments. The first method is based on Coupled Sparse Non-negative Matrix Factorization (CSNMF) of spectral representations and their corresponding class activity annotations. The second method is based on Multi-class Random Forest (MRF) classification of time-frequency patches. We compare the performance of the two methods on a recently published dataset TUT Sound Events 2016 containing data from home and residential area environments. Both methods show comparable performance to the baseline system proposed for DCASE 2016 Challenge on the development dataset with MRF outperforming the baseline on the evaluation dataset.

Index Terms— Acoustic event detection, random forest classifier, non-negative matrix factorization, sparse representation

1. INTRODUCTION

Acoustic Event Detection (AED) is an important task in machine listening. It aims to automatically recognise, label, and estimate the position in time in a continuous audio signal of meaningful sounds, referred to as acoustic events. There exists a number of real-world applications for AED such as home-care [1], surveillance [2], multimedia retrieval [3], urban traffic control [4], to name just a few. The task of AED can be broadly classified into *monophonic* and *polyphonic* detection. Monophonic detection, which has been the major area of research in this field, aims to recognize only one prominent event at a time [5, 6, 7]. However, in real-life environments multiple events occur at the same time making the task challenging. Polyphonic detection aims to identify these several overlapping events at the same time.

Several solutions have been proposed for polyphonic AED. Some approaches were strongly inspired by speech recognition systems, using mel frequency cepstral coefficients (MFCCs) with Gaussian Mixture Models (GMMs) combined with Hidden Markov Models (HMM) [8, 9]. Another popular technique for AED is matrix factorization of time-frequency spectra, especially Non-negative Matrix Factorization (NMF) [10]. NMF has been used to extract dictionaries for each acoustic event class in a supervised manner using isolated sounds [11, 12]. This approach served as a baseline for the Event Detection - Office Synthetic subtask of DCASE2013 Challenge [13]. In the same challenge, the best performance on polyphonic AED was achieved by examplar-based NMF decomposition followed by HMM postprocessing [14]. Another system used NMF to separate sound into different tracks and then detected events in each track separately assuming prior knowledge of the number of overlapping sources [15]. Probabilistic Latent Component Analysis (PLCA), the probabilistic counterpart of NMF, was also used for polyphonic AED using isolated sounds as training data [16].

Recently, several approaches for polyphonic AED that learn models directly from the mixture of sounds have been proposed. NMF applied to annotated overlapping events was used directly on the mixture of sounds, without the need to learn from isolated samples [17, 18]. Feed-forward deep neural networks (FNNs) trained on mixtures of sounds for multi-label AED achieved better performance than NMF [19]. Recurrent neural networks (RNNs) using bidirectional long short-term memory (BLSTM), which can directly model the sequential information of audio, were reported to outperform FNNs on the same dataset [20]. Despite their successes DNNs have several drawbacks. They are computationally complex and rely on huge amounts of data, hence data augmentation is often necessary to achieve better results [20]. Moreover, it is often difficult to interpret, what features does a DNN learn. On the contrary, methods such as NMF or multi-class classification are less computationally complex, and, in the case of NMF, may offer interpretable dictionaries.

In this paper, we propose two methods for polyphonic AED. The first method is inspired by promising results of coupled matrix factorization in [18]. We explore this idea by modifying the learning algorithm to explicitly sparse NMF. The second method is based on a multi-class random forest classification [21]. The random forest classifier has proved efficient for environmental sound classification and for monophonic AED [22, 6]. Therefore, we explore its application to polyphonic AED. As feature input for both methods we chose 2D time-frequency patches, which have proven effective for standard NMF [7]. We investigate the influence of the size of the patch. We compare our results on the TUT Sound Events 2016 introduced for the DCASE2016 Challenge [23].

This paper is organised as follows: Section 2 presents the details of training and testing procedures of the two methods. The experimental setup and the results of the evaluation are shown in Section 3. Section 4 contains the discussion of the results and conclusions. Conclusions and future work are presented in Section 5.

The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement no 642685 MacSeNet. MDP is also partly supported by EPSRC grant EP/NO14111/1

2. METHOD

We propose two methods for polyphonic AED and compare them on the development set of Task 3 of the DCASE2016 Challenge. The first method is based on sparse dictionary learning using non-negative matrix factorization (NMF) on 2D spectral patches coupled with class annotations. The second is based on a simple multi class random forest classification of 2D spectral patches. Both methods classify directly the mixture of events instead of building separate models for each sound event. The output of each method per frame is directly a set of multiple labels class activity. In order to make a fair comparison between the methods, we use the same preprocessing and post processing for both approaches. Both methods are implemented using *python*. For audio processing we used *librosa* [24] and for machine learning *scikit-learn* [25] libraries.

2.1. Preprocessing

We pre-process the data to reduce the dimensionality but at the same time remain meaningful representation appropriate for environmental sounds. Therefore, as a feature representation we choose to use the perceptually motivated mel scale [26]. We extract mel-spectrograms with 40 components, using a window size of 23 ms, hop size of the same duration and sampling frequency of 44.1 kHz. 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 spectral patches, also known as *shingling*, which has proven to be a discriminative feature for environmental audio classification [22]. We investigate the size of patches as it may differ depending on the characteristics of the sounds that we are trying to detect. Finally, the patches are normalised to account for intensity level difference among different occurrences of the events.

2.2. Coupled Sparse Non-negative Matrix Factorization

Coupled Sparse Non-negative Matrix Factorization (CSNMF) is inspired by the approach by Dikmen et al. [17]. The system presented by the authors uses coupled matrix factorization to learn dictionaries based on spectral representation of signals and the corresponding labels. In our method, we use sparse NMF to learn the coupled dictionaries in an analogical way.

The aim of a standard NMF is to find a low-rank representation of a matrix $\mathbf{V}$ by approximating it as a product of a non-negative dictionary $\mathbf{W}$ and its non-negative activation matrix $\mathbf{H}$, so that:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\mathbf{H} \tag{1}$$

where $\mathbf{V} \in R_+^{F\times N}$, $\mathbf{W} \in R_+^{F\times K}$ and $\mathbf{H} \in R_+^{K\times N}$.

In a coupled matrix factorization problem we are given two different matrices $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$, which we want to decompose into non-negative dictionaries $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ which share a common activation matrix $\mathbf{H}$. For polyphonic AED, $\mathbf{V}^{(1)}$ is a spectral representation of the signal of size $F \times N$, $\mathbf{V}^{(2)}$ is a binary matrix of class activation of size $E \times N$. F is the number of frequency bins, N number of frames and E number of classes of events. The dictionaries $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are found by minimizing the following objective function:

$$\eta_1 D^{(1)}(\mathbf{V}^{(1)}||\mathbf{W}^{(1)}\mathbf{H}) + \eta_2 D^{(2)}(\mathbf{V}^{(2)}||\mathbf{W}^{(2)}\mathbf{H}) + \lambda||\mathbf{H}||_1 \tag{2}$$

where $D(\mathbf{V}||\hat{\mathbf{V}})$ is chosen to be Kullbeck-Leibler (KL) divergence between the data V and the approximation $\hat{\mathbf{V}}$ and λ is a regularization parameter that penalizes over the l_1-norm, which induces sparsity on activation matrix $\mathbf{H}$. To facilitate computation the weighting parameters are chosen to be equal, i.e $\eta_1 = \eta_2 = 1$. We choose 60 bases for NMF, the number selected empirically.

The authors of [17] used an estimator based on maximum marginal likelihood (MMLE), which was shown to return sparse solutions. In order to investigate the influence of sparsity on the performance of the algorithm, we introduce the sparse regularisation explicitly and minimise the objective function in (2) using the multiplicative update rule for NMF with a sparsity constraint λ on the activation matrix [10, 27]:

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

where $\mathbf{V}$ is a concatenation of $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$, $\mathbf{W}$ is a concatenation of $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, $\mathbf{H}$ the common activation matrix, λ the sparsity regularizer and 1 is a matrix of ones of the size of V. $A \odot B$ denotes a Hadamard product of two matrices, A/B - Hadamard division and other multiplications are matrix multiplications.

Having learnt the dictionaries $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, in the testing phase we obtain an activation matrix $\mathbf{H}_{test}$ based on spectral representation of the test data, $\mathbf{V}^{(1)}$, and its dictionary $\mathbf{W}^{(1)}$. Next, we obtain the class activity matrix, $\mathbf{V}^{(2)}$, by multiplying the activation matrix $\mathbf{H}_{test}$ with the label dictionary $\mathbf{W}^{(2)}$. More details can be found in [17]. The estimated class activity matrix needs to be binarised using an arbitrary threshold to show the presence or absence of the sound.

2.3. Multi-class random forest classification

The multi-class random forest classification (MRF) method is based on classification of spectro-temporal patches using random forest classifier of 500 trees. This combination of features and classifier has proved to perform well on environmental sound classification task [22]. As the authors of [22] classify single events only, we modify the method to perform polyphonic AED. For E events in the dataset, we model all their possible combinations, i.e. we construct $M = 2^E$ classes. That means, that the set M of possible classes is a Cartesian product of $\{0, 1\}^E$. We classify each patch as belonging to one of the M product classes and concatenate all the estimates to form a class activation matrix. We can only model combinations of sounds already seen in the training set. Hence, any new combination of audio events will not be recognised correctly.

2.4. Postprocessing

We post-process the annotation matrices obtained by both methods using the baseline approach for DCASE2016, i.e. discarding events shorter than 100 ms and removing gaps shorter than 100 ms between the events.

3. EXPERIMENTS

3.1. Experimental setup and metrics

The two methods, CNMF and MRF, are tested on the TUT Sound Events 2016 dataset provided as a development set for Task 3 of

DCASE2016 Challenge [23]. The dataset consists of two everyday environments: one outdoor environment which is residential area with 7 classes, and one indoor environment, home with 11 classes. Table 1 shows the list of classes and the number of instances for both acoustic scenes. We can clearly see that the dataset is unbalanced, especially the residential area acoustic scene. Two classes, namely "bird singing" and "cars passing by", account for 74% of all the class instances. The home acoustic scene dataset is more balanced, but the two most appearing classes, i.e. "dishes", and "object impact", account for 42% of all instances.

Residential area		**Home**	
Event class	instances	Event class	instances
object (banging)	23	(object) rustling	60
bird singing	271	(object) snapping	57
car passing by	108	cupboard	40
children shouting	31	cutlery	76
people speaking	52	dishes	151
people walking	44	drawer	51
wind blowing	30	glass jingling	36
		object impact	250
		people walking	54
		washing dishes	84
		water tap running	47

Table 1: TUT Sound Events 2016: event classes and number of instances

As a metric for evaluating the methods we chose the official measure proposed for DCASE2016 Challenge, that is segment-based F-score given by the formula:

$$F = \frac{2P \cdot R}{P + R}, P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \qquad (4)$$

where TP is the number of true positives, FP - false positives, and FN - false negatives.

The second metric that we report is Error Rate (ER), proposed for DCASE2016, which measures the amount of errors in terms of a number of insertions $I(k)$, deletions $D(k)$ and substitutions $S(k)$ in a segment k [23]. The Error Rate is then calculated by integrating segment-wise counts over the total number of segments K, with $N(k)$ being the number of active ground truth events in segment k:

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \qquad (5)$$

Both F-score and Error Rate are calculated in 1 second segments.

3.2. Results

3.2.1. Binarization of activity matrix

The activation matrix that we get from CSNMF method needs to be binarized with a certain threshold in order to determine the presence of events. To determine the threshold value, we search among a number of threshold values to find the best balance between the F-score and Error Rate. For residential area environment the threshold is chosen to be of 30% of the maximum activation value, whereas for home environment - 20% of the maximum activation value.

3.2.2. Temporal context and sparsity

Table 2 and 3 show the influence of the sparsity regularizer and size of the 2D time-frequency patches on the performance of the proposed CSNMF method in residential area and home environment respectively. For the residential area we can see that adding some temporal context improves the F-score. The best result is achieved for 4 concatenated frames. Longer context is not beneficial for the method. Similarly, enforcing sparsity improves the results with the highest performance for $\lambda = 0.3$. However, for home environment the best performance is achieved with high sparsity, $\lambda = 0.5$, but using a single time frame as an input.

λ	0	0.1	0.2	0.3	0.5	1
1 frame	22.8%	30.5%	31.9%	19.3%	25.4%	23.5%
	2.45	1.37	1.13	1.03	1.11	0.95
4 frames	32.7%	33.9%	32.9%	**35.8%**	29.8%	11.7%
	1.46	1.09	0.95	**0.86**	0.95	1.03
6 frames	36.0%	34.2%	36.7%	35.1%	31.5%	16.3%
	0.96	1.01	0.91	0.96	0.92	1.05
8 frames	32.5%	42.0%	31.0%	32.6%	27.6%	14.6%
	1.21	0.95	1.02	0.96	1.00	1.06
10 frames	36.9%	43.2%	29.6%	23.6%	23.3%	15.8%
	0.96	0.96	1.00	1.07	1.04	1.09
12 frames	37.4%	37.5%	34.6%	21.7%	16.8%	12.6%
	0.97	0.95	1.04	1.04	1.10	1.09

Table 2: Residential area environment: F-score (%) and Error Rate (ratio) for different values of sparsity regularizer λ and temporal context (number of concatenated frames) using CSNMF

λ	0	0.1	0.2	0.3	0.5	1
1 frame	8.3%	8.0%	7.1%	7.5%	**11.7%**	11.4%
	5.72	3.18	1.73	1.79	**1.25**	1.32
4 frames	6.0%	7.4%	6.8%	10.0%	7.9%	11.6%
	3.13	1.83	1.36	1.41	1.35	1.35
6 frames	8.4%	9.3%	6.3%	6.7%	9.3%	10.1%
	1.77	1.58	1.40	1.51	1.48	1.65
8 frames	7.9%	10.0%	7.9%	6.5%	6.6%	7.6%
	1.62	1.48	1.39	1.40	1.60	1.82
10 frames	4.9%	6.6%	7.7%	8.6%	6.8%	7.0%
	1.52	1.55	1.45	1.56	1.74	1.94
12 frames	6.6%	6.4%	5.8%	6.9%	5.1%	10.0%
	1.51	1.37	1.52	1.72	1.78	2.03

Table 3: Home environment: F-score (%) and Error Rate (ratio) for different values of sparsity regularizer λ and temporal context (number of concatenated frames) using CSNMF

Table 4 shows the influence of the temporal context for the second proposed method, i.e. MRF classification. The best F-score and Error Rate for residential area environment is achieved for 2D time-frequency patches of 4 concatenated frames. For home environment the best result is achieved for 1 frame only.

3.2.3. Method comparison

The results on the development dataset for each context for the baseline, our Coupled Sparse Non-negative Matrix Factorization

	Residential area		**Home**	
	F	ER	F	ER
1 frame	36.8%	0.84	10.3%	**0.97**
4 frames	44.7%	**0.83**	14.9%	0.98
6 frames	**45.0%**	0.85	16.8%	0.98
8 frames	44.2%	0.84	**17.4%**	0.98
10 frames	44.1%	0.84	17.3%	0.99
12 frames	43.7%	0.83	16.2%	0.99

Table 4: F-score and Error Rate for different temporal context (number of concatenated frames) using Multi-class Random Forest (MRF) for home and residential area environments

(CSNMF) approach and our Multi-class Random Forest (MRF) approach in terms of F-score and Error Rate are shown in Table 5. As a baseline we chose the system provided for the DCASE2016 Challenge, which is based on MFCC acoustic features and GMM classifier [23]. On the development set, the MRF classifier outperforms the baseline in residential area environment and achieves comparable average results. CSNMF achieves similar F-score as the baseline but with a higher Error Rate. The results on the evaluation dataset are shown in Table 6. MRF outperformed the baseline and CNMF in both F-score and Error Rate.

	Residential area		**Home**		**Average**	
System	F	ER	F	ER	F	ER
baseline	31.5%	0.86	15.9%	**0.96**	23.7%	**0.91**
CSNMF	35.8%	0.86	11.7%	1.25	23.8%	1.06
MRF	**44.7%**	**0.83**	**17.4%**	0.98	**31.1%**	**0.91**

Table 5: Development dataset: Results for each context for the baseline system, Coupled Sparse Non-negative Matrix Factorization (CSNMF) approach and Multi-class Random Forest (MRF) approach in terms of F-score and Error Rate

Evaluation dataset		
System	F	ER
baseline	34.3%	0.88
CSNMF	29.2%	1.07
MRF	**44.1%**	**0.82**

Table 6: Evaluation dataset: Average results for the baseline system, Coupled Sparse Non-negative Matrix Factorization (CSNMF) approach and Multi-class Random Forest (MRF) approach in terms of F-score and Error Rate

4. DISCUSSION

We observe much lower performance of each of the compared methods on the home environment of the development dataset. This may be due to higher level of polyphony.

We have investigated the influence of the temporal context, i.e. number of concatenated frames, on the performance of the systems. Taking CSNMF method into consideration, introducing 2D spectral patches was beneficial for the residential area acoustic scene, however, it did not improve the performance for the home acoustic scene. That may be due to the fact that in the home environment we experience more impact sounds, such as “object impact”, “cupboard”, “object snapping”. Therefore, longer temporal context may blur the information contained in a single frame. On the contrary, residential area contains many sounds with long characteristics, such as “car passing by”, “bird singing”, which are at the same time the most frequent ones in the dataset.

We also investigated the influence of inducing explicit sparsity on matrix factorization, which has lead to better performance. Higher sparsity needs to be imposed for shorter temporal context.

An interesting discussion is brought up by analysing the results on the residential area environment of the evaluation dataset. As seen in Table 7, the algorithms, especially MRF, in fact recognize two classes very well, i.e. “bird singing” and “car passing by”, appearing much more often in the dataset than the others. Nevertheless, the average F-score of MRF is 9.9 percentage points higher than the baseline. Such a result shows the necessity of either evaluating the algorithms on a balanced dataset, where the events are distributed more evenly or using class-wise metrics to compare the performance of the algorithms.

		MRF		**CSNMF**	
Event label	Nref	F	ER	F	ER
(object) banging	11	0.0%	1.00	0.0%	1.27
bird singing	413	54.7%	1.15	56.3%	1.54
car passing by	213	68.6%	0.65	16.1%	0.98
children shouting	15	0.0%	1.00	0.0%	1.13
people speaking	57	0.0%	1.14	16.3%	2.16
people walking	146	0.0%	1.00	10.1%	1.46
wind blowing	48	0.0%	1.00	4.1%	0.98

Table 7: Residential area environment: F-score, Error Rate and number of references (Nref) in the dataset for each class using Multi-class Random Forest (MRF) and Coupled Sparse NMF (CSNMF)

5. CONCLUSION

In this paper we presented two methods for polyphonic AED in real environments. Both of the presented approaches achieve similar segment-wise F-score and Error Rate to the DCASE2016 baseline system on the TUT Sound Events 2016 dataset. The proposed MRF classification of spectral patches outperformed significantly the baseline on evaluation dataset despite its obvious drawbacks, such as incapability of recognizing a combination of sounds that was not present in the training dataset. Therefore, in the future we will investigate the possibilities of generating new combinations of sounds by allowing for multi-label classification. Moreover, spectro-temporal patches retrieved by shingling are a straightforward way for modelling temporal context, which did not improve detection on home environment. Therefore, we will investigate more complex time-frequency structure descriptors, such as the scattering transform. Additionally, we will investigate multiresolution approaches to model both short and long acoustic events.

6. REFERENCES

[1] P. V. Hengel and J. Anemüller, “Audio event detection for in-home care,” in *Proceedings of International Conference on Acoustics (NAG-DAGA)*, 2009, pp. 618–620.

[2] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, 2014.

[3] M. Bugalho, J. Portêlo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009)*, 2009, pp. 1151–1154.

[4] F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci, and P. Desii, "A real-time siren detector to improve safety of guide in traffic environment," *Proceedings of the 16th European Signal Processing Conference (EUSIPCO 2008)*, 2008.

[5] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *Proceedings of the 18th European Signal Processing Conference (EUSIPCO 2010)*, 2010, pp. 1267–1271.

[6] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 20–31, 2015.

[7] 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.

[8] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013.

[9] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," 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/DHV.pdf

[10] 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.

[11] 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 *Proceedings of the 21st European Signal Processing Conference (EUSIPCO 2013)*, vol. 47, 2013, pp. 552–567.

[12] A. Dessein, A. Cont, and G. Lemaitre, *Real-time detection of overlapping sound events with non-negative matrix factorization*, F. Nielsen and R. Bhatia, Eds., 2012.

[13] D. Stowell, D. Giannoulis, and E. Benetos, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17 (10), pp. 1733–1746, 2015.

[14] 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

[15] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, 2013, pp. 8677–8681.

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

[17] 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.

[18] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 2015, pp. 151–155.

[19] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2015)*, 2015, pp. 2551–2555.

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

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

[22] 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.

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

[24] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference (SciPy 2015)*, K. Huff and J. Bergstra, Eds., 2015, pp. 18–25.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, 1937.

[27] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, no. 4, 2004, pp. 2529–2533.