ON THE JOINT USE OF NMF AND CLASSIFICATION FOR OVERLAPPING ACOUSTIC EVENT DETECTION

Panagiotis Giannoulis[*,†], *Gerasimos Potamianos*[§,†], *Petros Maragos*[*,†]

[*] School of ECE, National Technical University of Athens, 15773 Athens, Greece
[§] Department of ECE, University of Thessaly, 38221 Volos, Greece
[†] Athena Research and Innovation Center, 15125 Maroussi, Greece

ABSTRACT

In this paper, we investigate the performance of classifier-based non-negative matrix factorization (NMF) methods for detecting overlapping acoustic events. We provide evidence that the performance of classifier-based NMF systems deteriorates significantly in overlapped scenarios in case mixed observations are unavailable during training. To this end, we propose a K-means based method for artificial generation of mixed data. The method of Mixture of Local Dictionaries (MLD) is employed for the building of the NMF dictionary using both the isolated and artificially mixed data. Finally an SVM classifier is trained for each of the isolated and mixed event classes, using the corresponding MLD-NMF activations from the training set. The proposed system, tested on two experiments with a) synthetic and b) real events, outperforms the state-of-the-art classifier-based NMF system in the overlapped scenarios.

Index Terms— NMF, MLD, Mixed Data, Overlapping Acoustic Event Detection

1. INTRODUCTION

Acoustic event detection (AED) is a major part of the computational auditory analysis field, aiming to detect the time boundaries of meaningful sound events. With audio being a crucial modality in multimodal content, most common applications of AED include smart home environments, surveillance and security [1, 2], as well as multimedia database retrieval.

Several methods have been developed the last years for AED. In the case of isolated AED, traditional methods based on hidden Markov models (HMMs) in conjunction with conventional features (e.g., MFCCs) show satisfactory performance [3, 4]. Regarding the more challenging overlapped scenario, different approaches include temporally-constrained probabilistic analysis models [5], generalized Hough-transform based systems [6], HMM-based systems with multiple-path Viterbi decoding [7], non-negative matrix factorization [8], and multi-label deep neural networks. In particular, the latter have shown good performance by modeling overlapping events in a natural way [9, 10].

NMF-based approaches constitute a popular choice for AED, and especially when it comes to overlapping scenarios, due to their natural relation with the source separation task and their ability to detect multiple events occurring simultaneously. NMF-related methods can be separated in those that exploit the NMF activations directly to perform event detection [8, 11], and in those that employ a classifier trained on these activations [12, 13]. Based on the fact that NMF-based approaches can benefit from the creation of a Mixture of Local Dictionaries (MLD) [14], in [15] the authors propose a classifier-based NMF system using MLDs for improved detection performance.

In our paper we investigate the performance of state-of-the-art NMF approaches under overlapped conditions. We provide evidence that the performance of the, so far, classifier based NMF methods degrades significantly in overlapped scenarios, mainly because the training phase considers activations only from isolated data. To alleviate this problem, we propose the generation of mixed observations using the isolated ones available, and subsequently their incorporation in the training data. For the artificial mixing procedure, we use a K-means based method for each pair of events. The MLD dictionary is built using the new training set, and SVM classifiers are trained for each of the isolated and mixed events using the corresponding activations. Our method is tested in two experiments using a) synthetic and b) real event instances and shows significant improvement over the state-of-the-art classifier based method in the overlapping scenarios.

The remainder of the paper is organized as follows: Section 2 presents and discusses the drawbacks of the two NMF-based alternatives that are compared with our system; Section 3 describes the artificial generation of mixed data and the outline of the proposed method; Section 4 reviews the experimental framework and reports our results; and, finally, Section 5 concludes the paper.

2. EXISTING NMF-BASED METHODS FOR AED

We will present briefly two popular methods for NMF-based AED. The first can be considered as the baseline, as it is the

This work has been partially funded by the BabyRobot project, supported by the EU Horizon 2020 Programme under grant 687831.

simplest one: Sparse-NMF with thresholding. The second is a classifier-based MLD-NMF method presented in [13, 15]. We will discuss the drawbacks of these two methods for isolated/overlapped acoustic event detection.

2.1. Sparse-NMF approach

The application of sparse-NMF for isolated and overlapping AED is based on the idea of linear decomposition of events into spectral atoms. Given non-negative features with approximate linearity (e.g. filterbank energies), a test event will be decomposed into atoms of observed event(s).

NMF is a linear non-negative approximate factorization of the observed feature matrix, and it is formulated as follows: Given a non-negative matrix $\mathbf{V} \in \Re^{\geq 0, M \times N}$, the goal is to approximate $\mathbf{V}$ with the product: $\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$, where $\mathbf{W} \in \Re^{\geq 0, M \times R}$ denotes the non-negative dictionary matrix, and $\mathbf{H} \in \Re^{\geq 0, R \times N}$ represents the non-negative activation matrix. Minimization of a suitable error cost function $D(\mathbf{V}||\mathbf{WH})$ results in iterative estimation of $\mathbf{W}$ and $\mathbf{H}$ [16].

For detection, assuming a given dictionary $\mathbf{W}$ that contains atoms of the various classes of interest, the estimated $\mathbf{H}$ provides activations of each class through time. It is shown that the sparse-NMF which imposes sparsity on the matrix $\mathbf{H}$, performs better for the detection task. Sparse-NMF minimizes the following objective: $D(\mathbf{V}||\mathbf{WH}) + \lambda\|\mathbf{H}\|_1$, with D denoting the generalized KL-divergence between $\mathbf{V}$ and $\mathbf{WH}$, and parameter λ controlling the trade-off between sparseness on $\mathbf{H}$ and accurate reconstruction of $\mathbf{V}$.

The method is used in this paper as a baseline. Regarding the building of the dictionary, using training data consisting of isolated event instances, a sufficient number of atoms is extracted and stored in the dictionary for each class of interest, resulting in the total dictionary matrix $\mathbf{W}$. Then in the detection step, a simple thresholding on the activations of matrix $\mathbf{H}$ decides for the existence of each event in each frame.

We can note two main disadvantages in this traditional method. The first is that the threshold-based decision in the detection step cannot be considered as the best choice in terms of robustness. The second and more important, is that, as pointed out in [14], the convex cones created by the bases of the sub-dictionaries of the different classes may often overlap between each other. This means that new observations that fall in the overlapped regions can be reconstructed with many different ways (unstable activations) which can result in failure of classification (e.g. false alarms).

2.2. SVM-based NMF approach with MLD dictionary

This method essentially refers to the core system of the works in [13, 15]. This system attempts to overcome the drawbacks of the aforementioned traditional sparse-NMF method by employing an MLD dictionary framework and an SVM classifier for the final detection step. The MLD-based dictionary generation eliminates overlaps between convex cones, and produces more stable activations which are used for the training of robust SVM classifiers. As shown in the flow diagram in Fig. 1 (black schemes), the method consists of two main parts; dictionary learning and classifier training.

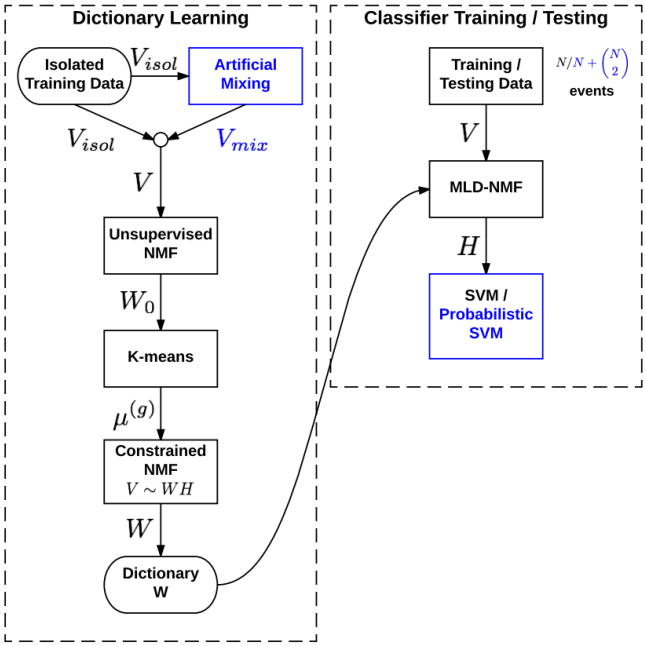

Fig. 1: Block-diagram of the proposed AED method.

Dictionary learning

In dictionary learning, the feature matrix $\mathbf{V}$ containing all training data is decomposed into an initial basis matrix $\mathbf{W}_0$ by basic unsupervised NMF. Next, by applying K-means to $\mathbf{W}_0$, G centroids $\mu^{(g)}$ are obtained, with $g \in \{1, ..., G\}$ denoting the centroid's index. The final MLD dictionary $\mathbf{W}$ consists of G sub-groups (of K_g bases each) which model acoustic atoms $\mathbf{W} = [\mathbf{W}^{(1)}...\mathbf{W}^{(G)}]$. The MLD dictionary is learned by minimizing the following objective:

$$D(\mathbf{V}||\mathbf{WH}) + \eta \sum_g D(\mu^{(g)} \| \mathbf{W}^{(G)}) + \lambda \sum_t \Omega(\mathbf{h}_t)$$

where $\mathbf{h}_t$ denotes the column vector of $\mathbf{H}$ at time frame t. The second term is a constraint which makes bases of sub-groups to be similar with $\mu^{(g)}$, so that the resulting convex cones are compact. The third term preserves group-sparsity in the solution.

Classifier training

For each class considered, an activation matrix $\mathbf{H}_i$ is extracted from its corresponding training spectrogram $\mathbf{V}_i$ by MLD based NMF with the global dictionary $\mathbf{W}$. Then the column vectors $\mathbf{h}_{t(i)}$ of $\mathbf{H}_i$ at each time frame t are used as feature vectors to train a linear SVM classifier. A multi-class SVM is trained using the one-against-all approach.

This method seems to solve the problems of the traditional sparse-NMF approach in the isolated AED case. Although, we must remark one possible drawback in the case of overlapping scenarios: The classifiers are trained for each class of

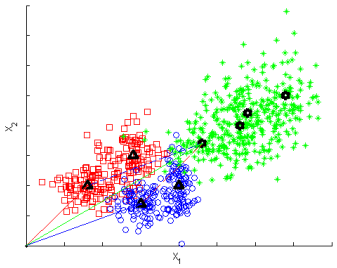

Fig. 2: Generation of mixed data (green) from a pair of isolated events (blue and red). Toy example, with two features "x_1" and "x_2".

interest using its corresponding isolated data. This makes the classifier vulnerable in the presence of unseen mixed data. An observation of a mixed event containing classes i and j will not necessarily be classified correctly by both the classifiers of i-th and j-th event.

3. PROPOSED METHOD

Our method attempts to solve the deficiency of the previous method in overlapped scenarios, by considering mixed data in the training and testing stages. The block-diagram of the proposed method is depicted in Fig. 1 (black and blue schemes).

3.1. Dictionary learning

Our scope is to include mixed data in the dictionary learning procedure. Considering the difficulty of having enough amount of mixed data available, we propose a method for artificial generation of mixed data. Assuming linearity of features, the method acts in the feature and not in the signal domain. The basic idea is shown in Fig. 2. In order to create representative observations of the mixed data, we try to combine (sum) representative observations from each of the two events considered.

Given a number of centroids C and a percentage α, we first perform K-means clustering with C clusters in the feature space of each event. Then $\alpha\%$ from the samples of each cluster are selected. Finally we consider all the combinations (addition) between the selected samples of the two classes.

After mixed data generation, both isolated and mixed data are used as input for the MLD dictionary learning procedure. In this way, bases created in the final dictionary may correspond to overlapped events too.

3.2. Classifier training

In the classifier training stage, instead of training N classifiers (N is the number of events), we train $N + \binom{N}{2}$. Also as we are modeling all the possible events (isolated and mixed), we train linear probabilistic SVMs and in the testing stage we choose the event with the highest score for each frame.

4. EXPERIMENTS

4.1. Datasets and Experimental Framework

We perform our experiments on two datasets, with the one containing synthetic events and the other real events. In

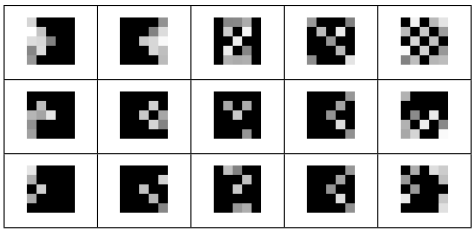

Fig. 3: Different instances for each of the 5 synthetic events. Horizontal axis corresponds to time and vertical to frequency.

the case of the synthetic event dataset, we generated artificial spectral patches for 5 synthetic events, while in the real event case, we extracted spectral patches from 5 real events contained in the database designed for the Task 2 of the DCASE'16 challenge (office-related events; drawer, phone, keys, speech, doorslam).

In both datasets, the performance of different methods is evaluated in both isolated and overlapped scenarios. In the isolated case, testing sequences of isolated spectral patches are created, whereas in the overlapped case, sequences of mixed spectral patches are generated. A mixed spectral patch results from the superposition of two isolated spectral patches from the corresponding testing dataset. Regarding the spectral patch extraction, in the case of synthetic events, we generate 5x5 spectral patches with the following procedure: The spectral patches of each event are characterized by a particular pattern which is slightly varying its structure in the different instances (see Fig. 3). To introduce variability, each time some of the active "tiles" of the the pattern can be missing (up to 5), while the active "tiles" take random positive values in the $[0.5, 1]$ interval. Random noise is also added after the generation of each spectral patch. In the case of real events, spectral patches have dimension 100x10 and are composed of 100 Mel-filterbank energies in 100msec intervals (10 frames).

Finally, regarding the partition into training and testing sets, in the real event case, we partitioned the training data of DCASE'16 challenge, so that 80% of event recordings is used for training and the rest 20% for testing purposes. In the synthetic event case, we generated a small number of instances per event (30) for building the training set. For both databases, the testing sequences contain 1000 spectral patches for both isolated and overlapped scenarios. We should note, that in the way that we build our synthetic testing sequences, when overlap occurs, it occurs in the whole duration of spectral patches involved. In this way, our problem can be also considered as classification of spectral patches of acoustic events with temporal information.

4.2. Results

In Tables 1 and 2, the comparative results for the three different methods are presented in terms of Fscore, for both isolated and overlapped scenarios and under two different experimental setups, for the two event datasets. In the first

Table 1: Performance of the different systems for the synthetic data scenario in terms of Fscore (%).

Method	Local opt.			Global opt.		
	Isol	Overl	Avg	Isol	Overl	Avg
sparse-NMF	95.10	95.82	95.46	95.21	93.53	**94.37**
SVM&MLD-NMF	96.78	77.23	87.00	94.39	77.23	85.81
Proposed	96.42	94.80	**95.61**	92.30	91.16	91.73

Table 2: Performance of the different systems for the real data scenario in terms of Fscore (%).

Method	Local opt.			Global opt.		
	Isol	Overl	Avg	Isol	Overl	Avg
sparse-NMF	78.36	78.54	78.45	75.49	77.52	**76.51**
SVM&MLD-NMF	85.83	61.76	73.79	83.96	61.76	72.86
Proposed	85.79	74.49	**80.14**	82.00	68.86	75.43

setup (Local opt.), optimization of the various parameters of the methods is performed in each scenario separately, while in the second (Global opt.) optimization is performed only one time for the whole testing procedure. In fact, "Local opt." assumes prior knowledge of overlap existence.

In Table 1 we can draw three major conclusions: First of all, our proposed method clearly outperforms the state-of-the-art SVM&MLD-NMF based method in the overlapping scenarios, both in "local" and "global" setups achieving 77.16% and 61.18% relative error reductions correspondingly. In fact, SVM&MLD-NMF method's performance degrades significantly in the presence of mixed events. Next, we can observe that the performance of baseline sparse-NMF approach is stable across the different scenarios and setups, achieving also the best Fscore in the "global" optimization setup. We can say that in the case of quite simple and discriminable events this baseline is a good option for both isolated and overlapped scenarios. Finally, only our proposed method seems to be affected significantly by using global optimization instead of the local one. It seems that the parameter α that controls the amount of mixing data included in the training phase, has strong influence on the behavior of our method.

In Table 2, corresponding results for the real-event scenario are presented. Similarly to the synthetic case, we can again notice the big drop in the performance of SVM&MLD-NMF method when we move from the isolated to the overlapped scenario, as well as the superiority of the proposed method in the overlap case (33.29% and 18.57% relative error reduction in "local" and "global" setups respectively). Also, the baseline sparse-NMF method shows again stable performance across different scenarios. However, as expected, in this more challenging case of real events, both the SVM&MLD-NMF and proposed methods perform significantly better than the baseline in the isolated scenario. Finally, like before, among the three methods, our approach is affected the most by the switch from the "local" to the "global" optimization setup.

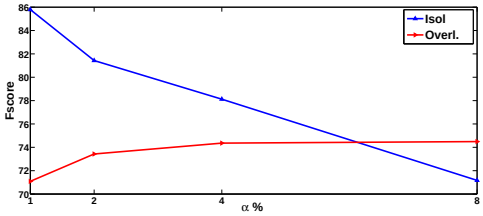

Fig. 4: Performance of the proposed method in both Isolated and Overlapped scenarios as the percentage α of mixing increases.

By summarizing the results, we can claim that the classifier based SVM&MLD-NMF approach outperforms the baseline sparse-NMF based one in the isolated event scenario. This is important, as the fact is that the isolated scenario is by far the most frequent under realistic conditions. However, if we want to test the system under more challenging overlapping conditions, the performance of the existing method deteriorates. Our proposed method, by incorporating mixed data in the training phase, succeeds to increase the performance significantly under overlapped conditions, and also provide better results in total. However there is one drawback: our method is strongly affected by the amount of mixed data employed for training. This is depicted also in Fig. 4, where the performance of the proposed method is shown for the real events dataset, for both the isolated and overlapped cases, as the mixing parameter α increases. As α increases, performance increases also in the overlapping case, but at the same time, decreases (with a higher rate) in the isolated case. With knowledge of the expected degree of overlap in our dataset, an optimal value of α could be chosen.

5. CONCLUSION

In this paper we investigated the performance of state-of-the-art NMF approaches for overlapping acoustic event detection. We provided evidence of degradation of the existing method's performance under highly overlapped conditions, and we proposed a new method which tries to alleviate this problem by employing a module for artificial generation of mixed data which are considered in the training phase. Probabilistic SVMs are also employed in the final classification step using all available classes (isolated and mixed).

Results obtained on experiments with synthetic and real events were promising, outperforming the existing method in overlapping scenarios while also preserving good performance in the isolated ones.

In future work, the design of a module able to identify the existence (or not) of overlap will be investigated, in order to increase the robustness of our system. Also alternative methods for artificial generation of mixed data will be considered.

6. REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005, pp. 1306–1309.

[2] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006, vol. 5, pp. V–V.

[3] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2375–2379.

[4] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with adaboost feature selection," in *Multimodal Technologies for Perception of Humans*. 2008, pp. 345–353, Springer.

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

[6] J. Dennis, H.D. Tran, and E.S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised Hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, 2013.

[7] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," *Proc. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (WASPAA)*, 2013.

[8] J.F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach to audio event detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[9] I. Choi, K. Kwon, S.H. Bae, and N.S. Kim, "DNN-based sound event detection with exemplar-based approach for noise reduction," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 16–19.

[10] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. International Joint Conference on Neural networks (IJCNN)*, 2015, pp. 1–7.

[11] P. Giannoulis, G. Potamianos, P. Maragos, and A. Katsamanis, "Improved dictionary selection and detection schemes in sparse-CNMF-based overlapping acoustic event detection," 2016.

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

[13] T. Komatsu, Y. Senda, and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2259–2263.

[14] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," in *IEEE Signal Processing Letters*, 2015, vol. 22, pp. 293–297.

[15] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," 2016.

[16] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.