IMPROVED DICTIONARY SELECTION AND DETECTION SCHEMES IN SPARSE-CNMF-BASED OVERLAPPING ACOUSTIC EVENT DETECTION

Panagiotis Giannoulis [1,3], *Gerasimos Potamianos* [2,3], *Petros Maragos* [1,3], *Athanasios Katsamanis* [1,3]

[1] School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2] Department of ECE, University of Thessaly, 38221 Volos, Greece
[3] Athena Research and Innovation Center, 15125 Maroussi, Greece

paniotis@central.ntua.gr, gpotam@ieee.org, maragos@cs.ntua.gr, nkatsam@cs.ntua.gr

ABSTRACT

In this paper, we investigate sparse convolutive non-negative matrix factorization (sparse-CNMF) for detecting overlapped acoustic events in single-channel audio, within the experimental framework of Task 2 of the DCASE'16 challenge. In particular, our main focus lies on the efficient creation of the dictionary, as well as on the detection scheme associated with the CNMF approach. Specifically, we propose a shift-invariant dictionary reduction method that outperforms standard CNMF-based dictionary building. Further, we develop a novel detection algorithm that combines information from the CNMF activation matrix and atom-based reconstruction residuals, achieving significant improvement over the conventional approach based on the activations alone. The resulting system, evaluated on the development set of Task 2 of the DCASE'16 Challenge, also achieves large gains over the traditional NMF baseline provided by the Challenge organizers.

Index Terms— Convolutive Non-Negative Matrix Factorization, Dictionary Building, Overlapped Acoustic Event Detection

1. INTRODUCTION

Acoustic event detection (AED) is a research topic that has been attracting increasing interest in the literature. Its main goal is the detection of "active" time intervals for each event present in an audio recording. In its general form, multiple acoustic events may occur simultaneously, making the task extremely challenging. Applications of AED include smart home environments, security and surveillance, and multimedia database retrieval, among others.

In the case of isolated AED, conventional detection and classification approaches, such as ones based on hidden Markov models (HMMs) in conjunction with traditional audio features (for example MFCCs) achieve satisfactory performance [1]. In the case of overlapped AED however, such methods need to be modified in order to allow multiple event detection. For example, in [2], multiple-path Viterbi decoding is employed to deal with the overlapping scenario. Other works for overlapping AED include multi-label deep neural networks [3], temporally-constrained probabilistic component analysis models, generalized Hough-transform based systems [4], and non-negative matrix factorization (NMF) [5].

Among these, NMF-based approaches and their variants have began to attract interest in the field of both isolated and overlapped AED in recent years. This is due to both their robustness and their natural ability to detect multiple events occurring simultaneously, as long as appropriate non-negative and linear representations of them are available. For example, in [6], a rather small dictionary is built automatically using sparse-CNMF, and subsequently the activations produced are used as input for HMM training for each class. Also in [5], using a large dictionary, NMF activations are directly exploited to perform detection for each event class.

In this paper, overlapping AED is performed on the Task 2 dataset of the DCASE'16 Challenge, containing 11 office-related events synthetically mixed in various conditions. The detection system proposed is based on the sparse-CNMF framework: Given a dictionary with spectral patches/atoms for each class, it determines the activations of each atom over time, thus allowing detection of overlapping events. The main contributions of the work lie in the investigation of methods for efficient dictionary building and in the design of a novel method for the final detection step. In particular, an efficient dictionary selection method based on shift-invariant similarity between atoms is proposed, achieving improved results compared to the standard automatic dictionary building of sparse-CNMF. Also, in the final detection step, a combination of activations with the reconstruction errors for each class is proposed. The results demonstrate remarkable improvement compared to the conventional approach of using activations alone, indicating the complementary information contained in the reconstruction errors.

The rest of the paper is organized as follows: Section 2 overviews the sparse-CNMF framework; Section 3 presents dictionary building for CNMF, including the proposed shift-invariant reduction approach; Section 5 covers the CNMF detection approaches considered; Sections 4 and 6 overview additional system components, such as background noise modeling, feature extraction, and post-processing; Sections 7 and 8 cover the experimental framework and results, and, finally, Section 9 concludes the paper.

2. SPARSE-CNMF FOR AED

The application of sparse-CNMF for overlapped AED is based on the idea of linear decomposition of events into spectral patches/atoms. Given the linearity of the features employed, mixtures of events will be mainly decomposed into atoms from the mixed classes, therefore indicating their presence. Non-negative features with approximate linearity are required: spectrograms and filterbank energies are typically used for this purpose.

NMF is a linear non-negative approximate factorization of the observed feature matrix. CNMF [7] is its convolutive extension, 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 convolutive sum:

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}} \tag{1}$$

where, in our case, $\mathbf{W}_t \in \Re^{\geq 0, M \times R}$ is the dictionary matrix at time step t, $\mathbf{H} \in \Re^{\geq 0, R \times N}$ the activation matrix, and T the length of each dictionary atom. The i-th column of $\mathbf{W}_t$ describes the i-th atom, t time steps after its beginning, and the $\overset{t\rightarrow}{\bullet}$ operator shifts the columns of its matrix argument t places to the right. The dictionary contains R atoms of size $M \times T$ each. If we denote with

$$\mathbf{\Lambda} = \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}} ,$$

the minimization of a suitable error cost function $D(\mathbf{V}||\mathbf{\Lambda})$ results in iterative estimation of $\mathbf{W}_t$ and $\mathbf{H}$ [7, 8].

For detection problems, given a dictionary $\mathbf{W}_t, t \in [0, T-1]$ containing patches/atoms for the various classes, the estimated $\mathbf{H}$ provides the activations of each class through time. Although CNMF produces activation patterns that tend to be sparse, in detection-related tasks sparsity of $\mathbf{H}$ becomes crucial. Sparse-CNMF, a variant of CNMF, minimizes the following objective:

$$G(\mathbf{V}||\mathbf{\Lambda}) = D(\mathbf{V}||\mathbf{\Lambda}) + \lambda \|\mathbf{H}\|_1 \quad . \tag{2}$$

Parameter λ controls the trade-off between sparseness on $\mathbf{H}$ and accurate reconstruction of $\mathbf{V}$. Depending on the cost function selected (KL-divergence, Euclidean distance) different updating equations result [9, 10].

3. DICTIONARY BUILDING

Dictionary building is a very important step in exemplar-based methods. Representative atoms from each class must be contained in the dictionary matrix, capable of reconstructing unseen data. Using training data of isolated event instances, a sufficient number of atoms is extracted and stored in the dictionary for each class:

$$\mathbf{W}_t = [\mathbf{W}_t^1 ... \mathbf{W}_t^C], \quad t \in [0, T-1] \quad , \tag{3}$$

where C is the number of classes. In the case of CNMF-based methods, due to increased computational complexity, we need to create a rather compact dictionary. We present two alternatives next.

3.1. CNMF-based

For each class, the training instances are concatenated to form the total $\mathbf{V}^i$ matrix. Then via sparse-CNMF, both matrices $\mathbf{W}_t^i$ and $\mathbf{H}^i$ are computed (as in [10]). Then, $\mathbf{W}_t^i \in \Re^{\geq 0, M \times R_i}$ is stored in the dictionary. The duration T of each atom and their total number R_i are predefined. By choosing to extract the same number of atoms R_i for each class, the total number of atoms will be $R = C \cdot R_i$.

3.2. Shift-invariant dictionary reduction

Here, we propose an alternative way for dictionary creation that selects a group of atoms from the original training data, instead of generating new ones as above. For each class, first, a large number of atoms is extracted from $\mathbf{V}^i$, using a sliding window of duration T (shifted by one feature frame at a time). Then only R_i of them are selected by uniformly sampling the space of available atoms. Uniform sampling aims at selecting different types of existing atoms based on a similarity measure, appropriate for CNMF. In our case, the similarity should be shift-invariant: i.e., two atoms are considered similar if the Euclidean distance between them or between their shifted versions is small.

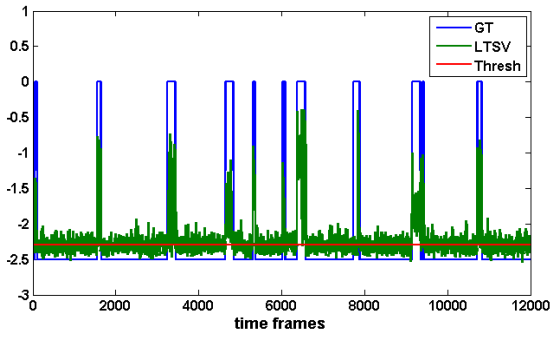


Figure 1: LTSV for background noise detection.

To compare two atoms in a shift-invariant way we first reshape them to vectors of size $M \cdot T$ in a row-wise manner. In this way, shift in time in the atoms results in the same shift for their reshaped vectors. Then similarity between atoms is measured as the Euclidean distance between the magnitudes of the Fourier transforms of their reshaped vectors, based on the well-known shift-invariant property:

$$|F\{w[t]\}| = |F\{w[t - to]\}| \tag{4}$$

where F denotes the DTFT transform. The different atoms are mapped to their Fourier-magnitude vectors, and the latter are sorted based on their Euclidean distance from their mean. Finally, R_i atoms are selected by uniformly sampling the sorted list.

4. BACKGROUND NOISE MODELING

In addition to the event modeling by incorporating representative atoms in the dictionary, background noise modeling is also necessary for a robust detection scheme. With the presence of background noise atoms in the dictionary, false alarm event activations are avoided in silent areas. Also, more reliable reconstruction is possible in active areas, assuming additive noise.

In our approach, following [5] we extract the background atoms from the observed data during the decoding (on-the-fly). The advantage of this approach is the adaptation of the background dictionary in the slightly different conditions existing each time.

However, instead of assuming background noise present in the beginning and end of the observed data as in [5], we attempt to extract background atoms from various areas of the signal, by employing the LTSV measure (Long-Term Signal Variability) described in [11]. This measure has been used successfully for the VAD task and is based on the fact that background noise usually exhibits smaller variability through time in its spectrum.

A frame is considered as noise if its LTSV value is lower than a fixed threshold T_L. As before, the shift-invariant dictionary reduction method is applied to noise areas occurred, to provide the background atoms. In Fig. 1 the LTSV values for an observed signal together with the ground-truth of event activations are shown.

5. DETECTION APPROACHES

As stated earlier, having created the dictionary matrix $\mathbf{W}_t$ the sparse-CNMF method takes as input the data matrix $\mathbf{V}$ and outputs the activation matrix $\mathbf{H}$ (following the approach in [9]). The final

event detection can occur by exploiting the information in the above matrices. We present two main approaches:

5.1. Activations only

Most of NMF-based approaches exploit the information in **H** directly [5] or indirectly [6]. In our method, activations in **H** are directly used for the detection of possible events. In particular, for each class, the activations are summed across all their atoms, for each frame, resulting in a new matrix $\mathbf{H}' \in \Re^{\geq 0, C \times N}$:

$$H'(i,t) = \sum_{atom \in i} H(atom, t) \quad (5)$$

where i is the class index ($i = 1, ..., C$). Then in time frame t a class is considered active if $H'(i,t) > T_c$. T_c is the activation threshold suitably selected. A post-processing step can also be employed in order to provide smooth activations. Finally, as activation refers to atoms, $T-1$ also frames after the detected activations are considered active.

5.2. Incorporation of Reconstruction Residuals

A complementary method of the previous one decides for the activation of an event not by thresholding the amplitude of $\mathbf{H}'$ but by measuring the KL-divergence reconstruction error if only the atoms of that event and of background noise are used. More specifically, the total reconstruction error of Sparse-CNMF for a time segment $seg = [t1, t2]$, is $D(\mathbf{V}_{seg}||\mathbf{\Lambda}_{seg})$ and the reconstruction error of the i-th event is $D(\mathbf{V}_{seg}||\mathbf{\Lambda}^{i,bg}_{seg})$, where:

$$\Lambda^{i,bg}_{seg} = \sum_{t=0}^{T-1} \mathbf{W}^{i,bg}_t \cdot \overset{t\rightarrow}{\mathbf{H}^{i,bg}_{seg}} \quad (6)$$

with $\mathbf{H}^{i,bg}$ denoting the part of **H** containing only the rows corresponding to atoms of the i-th class and of background noise.

We define the Residual Ratio of event-i as the ratio between the residual of event-i to the total residual using all the events:

$$RR(i,t) = \frac{D(\mathbf{V}_{seg}||\mathbf{\Lambda}^{i,bg}_{seg})}{D(\mathbf{V}_{seg}||\mathbf{\Lambda}_{seg})}, \quad t \in seg \quad (7)$$

For the **RR** computation, non-overlapping segments of 1 sec duration are used. Small Residual Ratio for the i-th event in a given segment means that large percentage of the reconstruction in that segment is achieved using only the i-th event (together with background noise).

In the first approach using activations only, the criterion for event detection is the amplitude of peaks in activation matrix **H**. In the residuals-based approach, the criterion is the accuracy of reconstruction using only atoms and activations of a particular event. In our final system we combine the above two approaches. The event-i is considered active in time frame t if both conditions hold:

$$H'(i,t) > T_c \quad \wedge \quad RR(i,t) < T_r \quad (8)$$

Thresholds T_c and T_r are chosen appropriately as explained later.

6. FEATURES, PARAMETERS AND POST-PROCESSING

Regarding the features, we have experimented with various feature sets that satisfy non-negativity and approximate linearity: Mel filterbank energies, Gammatone filterbank energies, DFT spectrogram, variable Q-Transform (VQT). The first three, are computed using frames of 30 msec with 10 msec shift. Regarding the Dictionary building, atoms of 200msec (17 frames) are used, and for the CNMF-framework the parameter λ was set to 0.7.

Concerning the various thresholds used, the threshold T_c for activations in $\mathbf{H}'$ matrix is computed as a percentage of the maximum peak of $\mathbf{H}'$. The threshold for residuals T_R is computed as a percentage of the minimum of **RR** matrix for a given segment. All the parameters were optimized using the Dev set (in our case it stands as Test set also).

Finally as a post-processing in the detection step, 1d-dilation is performed in each row of the $\mathbf{H}'$ matrix in order to broaden the intervals of high-peaked activations produced. In the case of the combined method, dilation is performed before the combination with the residuals approach. In the end, $T-1$ frames after each detected activation are also considered active.

7. EXPERIMENTAL FRAMEWORK

7.1. Database

We performed our experiments on the DCASE'16 Challenge database designed for Task 2: "Sound event detection in synthetic audio". This database contains recordings for 11 office-related audio events: clearing throat, coughing, door knock, door slam, drawer, human laughter, keyboard, keys, page turning, phone ringing and speech. The database consists of 3 parts: The Training set comprises of 20 isolated recordings for each event. The Dev set contains 18 recordings of synthetic mixtures with varying SNR levels, event density conditions and polyphony. Test set has similar structure with Dev set but is used only for the Challenge evaluation (ground-truth not available).

7.2. Setups

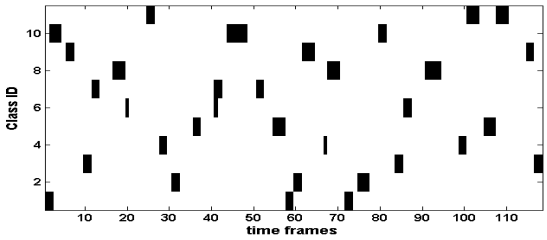
In this paper, we report experiments on the Dev set. However, due to its particularity of containing the same event instances with the Training set, we use two different setups.

The first setup is identical to the default setup of Task 2. One dictionary is built using all the isolated training data and then sparse-CNMF based detection is performed on each of the 18 Dev recordings.

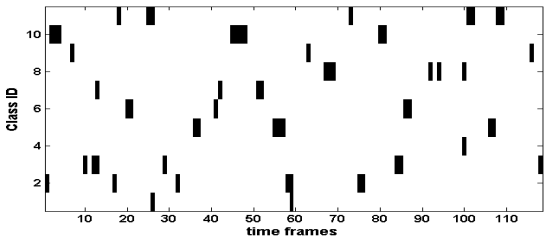
In the second setup, in order to test on unseen instances, we perform a 18-leave-one-out experiment. 18 dictionaries are built, one for each testing recording in Dev, by using each time all the training instances except of those contained in the given recording.

7.3. Metrics

Various metrics are considered in the Task 2, both event-based and frame-based. We will focus on the main metric of the challenge, the frame-based Total Error Rate (ER), defined as $ER = (I + D + S)/N$. I denotes insertions, D deletions, S substitutions and N is the total number of ground-truth events for a given frame. ER is computed in frames of 1 sec. Frame-based Fscore is also reported in our results.



(a)



(b)

Figure 2: Example of the event activations through time in (a) Ground-Truth (b) our system's output

8. RESULTS

Moving to the results, in Table 1 the results for the Challenge NMF-baseline and our system are compared using the default setup#1. Regarding the NMF-based baseline, it builds the dictionary using the training data, and extracts 20 atoms per class. Atoms have single-frame duration and are extracted from the variable-Q transform spectrogram (VQT, 60bins, 10 msec step). A post-processing stage applies median filtering to the output and allows up to 5 concurrent events.

Our system uses 200 atoms per event, of duration 200 msec each. The dictionary creation was performed employing our Shift-Invariant reduction method. It is obvious that our system (Activations only) clearly outperforms the baseline achieving a 69.6% relative improvement in terms of ER metric. It seems that the extraction of more atoms per class, combined with the incorporation of temporal structure in them under the CNMF-framework lead to major improvement.

In Table 2 we show our experimentation regarding different feature sets that can be used together with variations in their dimensionality and in dictionary size (#atoms per class). The results were conducted using the leave-one-out setup#2. We can observe that Mel filterbank energies achieve the best performance among the different feature sets. It seems that Mel filterbank is more appropriate for this group of events. Also from the results of the Mel features, we can see that increasing the dimensionality of the feature vector and the dictionary size can lead to slight improvements.

Comparison of different Dictionary building methods is shown in Table 3 using the setup#2. Also the same detection system was used in all cases (Mel-100-100). Our approach performs better than the standard CNMF-based dictionary building. This provides indication that accurate representation of event atoms (instead of approximate) is beneficial for the detection task, as long as we have a way to select the appropriate atoms.

In Table 4 the results of our systems using the two different detection approaches are depicted. We can observe that the system using the combination of activations and reconstruction residuals achieves a 20% relative error reduction compared to the system using activations only. This highlights the complementarity of the two approaches. The improvement is mainly due to the elimination of false activations exhibiting large peaks in $\mathbf{H}'$ matrix but having also large Residual Ratio.

Finally in Fig. 2 the output of our system is shown together with the ground-truth for a particular audio recording of Dev set.

Table 1: Performance of Baseline and our system for the detection task in the setup #1 experiment.

Setup #1		
Method	Fscore	ER
NMF-Baseline	0.42	0.79
Activations only	0.87	**0.24**

Table 2: Performance of different feature sets and dictionary sizes for the detection task in the setup #2 experiment.

Setup #2	
Features - dim.- size	ER
VQT - 545 - 200	0.88
Gamma - 150 - 200	0.86
Mel - 150 - 200	**0.79**
Mel - 150 - 100	0.82
Mel - 100 - 100	0.83
DFT - 545 - 100	0.83

Table 3: Performance of different dictionary building methods for the detection task in the setup #2 experiment.

Setup #2	
Method	ER
Sparse-CNMF	0.89
Shift-Inv. Reduction	**0.82**

Table 4: Performance of our systems for the detection task in the setup #2 experiment.

Setup #2		
Method	Fscore	ER
Activations only	0.43	0.79
Activations & Residuals	0.55	**0.63**

9. CONCLUSIONS

We presented a sparse-CNMF based system for overlapped audio event detection, employing an efficient dictionary building method and a novel detection approach. Special focus was also given to the background noise modeling and on the experimentation with different possible feature sets for the CNMF framework. Results presented for the Task 2 of DCASE16 challenge are promising and quite better than that of NMF-based baseline provided.

In future work, better ways to combine activation-based and residual-based approaches will be investigated. Also the performance of our system will be tested in more datasets relevant to overlapped AED.

10. REFERENCES

[1] 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.

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

[3] 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.

[4] J. Dennis, H. Tran, and E. 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.

[5] J. 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.

[6] C. Cotton and D. 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.

[7] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *International Conference on Independent Component Analysis and Signal Separation*, 2004, pp. 494–499.

[8] W. Wang, A. Cichocki, and J. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2858–2864, 2009.

[9] P. O'Grady and B. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Proc. 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2006, pp. 427–432.

[10] W. Wang, "Convolutive non-negative sparse coding," in *Proc. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 3681–3684.

[11] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.