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 overlapping 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 the detection scheme associated with the CNMF approach. Specifically, for the dictionary creation stage, we propose a shift-invariant method for its size reduction that outperforms standard CNMF-based dictionary building. Further, for detection, we develop a novel algorithm that combines information from the CNMF activation matrix and atom-based reconstruction residuals, achieving significant improvements over conventional detection based on activations alone. The resulting system, assisted by efficient background noise modeling, outperforms a traditional NMF baseline provided by the Challenge organizers, achieving a 24% relative reduction in the total error rate metric on the Challenge Task 2 test set.

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

1. INTRODUCTION

Acoustic event detection (AED) is a research topic that has attracted significant interest in the literature. Its main goal is the end-pointing and classification of each event present in an audio recording. In its general form, multiple acoustic events may occur simultaneously, making the task extremely challenging. Application areas of AED include, among others, smart home environments, surveillance and security, as well as multimedia database retrieval.

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 overlapping 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 [4], generalized Hough-transform based systems [5], and non-negative matrix factorization (NMF) [6].

Among these, NMF-based approaches and their variants have begun to attract interest in the field of both isolated and overlapping 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 [7], a rather small dictionary of events is automatically built using sparse-CNMF, and subsequently the activations produced are used as input for HMM training for each class. Also, in [6], 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 [8], consisting of single-channel audio that contains eleven 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 (acoustic event), 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 approach yields significant improvements over conventional detection employing activations alone, indicating the complementary information contained in the reconstruction errors.

The remainder 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 size reduction approach; Section 4 covers the CNMF detection approaches considered; Section 5 discusses additional system components, such as background noise modeling, feature extraction, and post-processing; Section 6 reviews the experimental framework and reports our results; and, finally, Section 7 concludes the paper.

2. SPARSE-CNMF FOR AED

The application of sparse-CNMF for overlapping 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. To accomplish this, 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 [9] is its convolutive extension, and

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

it is formulated as follows: Given a non-negative data feature matrix $\mathbf{V} \in \Re^{\geq 0, M \times N}$, where M denotes the feature vector size and N the available number of feature vectors, the goal is to approximate $\mathbf{V}$ by matrix $\mathbf{\Lambda}$, derived as a temporal convolutive sum of a "dictionary" and "activations", namely

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

where, operator $\overset{t\rightarrow}{\bullet}$ shifts the columns of its matrix argument t places to the right, $\mathbf{W}_t \in \Re^{\geq 0, M \times R}$ denotes the non-negative dictionary matrix at time step t, $\mathbf{H} \in \Re^{\geq 0, R \times N}$ represents the non-negative activation matrix, T is the number of time frames spanned by each dictionary atom, and R stands for the number of atoms in the dictionary. The i-th column of $\mathbf{W}_t$ describes the i-th atom, t time steps after its beginning. The dictionary thus contains R atoms of size $M \times T$ each. Minimization of a suitable error cost function $D(\mathbf{V} || \mathbf{\Lambda})$ results in the iterative estimation of $\mathbf{W}_t$ and $\mathbf{H}$ [9, 10].

For detection, assuming a given dictionary $\mathbf{W}_t$, $t \in [0, T-1]$, that contains atoms of the various classes of interest, 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. To achieve it, sparse-CNMF, a variant of CNMF, is often used, minimizing the following objective,

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

with parameter λ controlling the trade-off between sparseness on $\mathbf{H}$ and accurate reconstruction of $\mathbf{V}$ by $\mathbf{\Lambda}$. Depending on the cost function selected (KL-divergence, Euclidean distance), different updating equations result [11, 12].

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 consisting of isolated event instances, a sufficient number of atoms is extracted and stored in the dictionary for each class of interest, resulting to matrices

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

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. In the following, we present two alternatives for this task.

3.1. CNMF-based

For each class of interest, the training instances are concatenated to form its data matrix, $\mathbf{V}^{(i)}$. Then, via sparse-CNMF, matrices $\mathbf{W}_t^{(i)}$ and $\mathbf{H}^{(i)}$ are computed (as in [12]), and $\mathbf{W}_t^{(i)} \in \Re^{\geq 0, M \times R_i}$ stored in the dictionary. The duration, T, of each atom and their total number, R_i, are predefined. By extracting the same number of atoms for each class, their total number becomes $R = C \cdot R_i$.

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. For each class, first, a large number of atoms is extracted from its data matrix $\mathbf{V}^{(i)}$, 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 set of the resulting atoms, as explained next. The process aims at selecting different types of existing atoms based on a similarity measure, appropriate for CNMF. In our case, such similarity should be shift-invariant: i.e., two atoms are considered similar if the Euclidean distance between them, or between their temporally shifted versions, is small.

To achieve atom comparisons in a shift-invariant way, we first rearrange them into vectors of size $M \cdot T$, in a row-wise manner. This way, a time-shift of atoms results to shifts of their corresponding vectors. Then, atom similarity is measured as the Euclidean distance between the magnitudes of the Fourier transforms (DFTs) of the rearranged vectors, based on the well-known shift-invariant property of this transform. The available atoms are thus mapped to their Fourier-magnitude vectors, which are subsequently sorted based on their Euclidean distance from their mean. Finally, R_i atoms are selected by uniformly sampling the resulting sorted list.

The adopted sampling scheme represents a simple approach to desired dictionary size reduction. Alternatively, well-known clustering methods like k-means could also be used for the task.

4. DETECTION APPROACHES

As stated earlier, having created the dictionary matrix $\mathbf{W}_t$, sparse-CNMF accepts as input the data matrix $\mathbf{V}$, and outputs the desired activation matrix $\mathbf{H}$ (following the approach in [11]). The final event detection can occur by exploiting the information in the above matrices. We present two main approaches for accomplishing this.

4.1. Using activations only

Most of NMF-based approaches employ the information in $\mathbf{H}$ directly [6], or indirectly [7]. In our method, activations in $\mathbf{H}$ are directly used for detecting possible events. In particular, for each class, the activations are summed across all their atoms, for each frame, resulting in a new matrix $\mathbf{H}' \in \Re^{\geq 0, C \times N}$, with elements

$$H'(i, n) = \sum_{r \in \{i\}} H(r, n) \quad , \tag{4}$$

where i denotes the class ($i = 1, \cdots, C$), $\{i\}$ the set of row indices in $\mathbf{H}$ that correspond to the i-th event atoms, and $n \in \{1, \cdots, N\}$ the time frame. Then, at time n, a class is considered active if $H'(i,n) > \theta_H$, where θ_H is a suitably chosen activation threshold. A post-processing step can also be employed to yield smooth activations. Finally, as activation refers to atoms, $T-1$ additional frames following the detected activations are considered active.

4.2. Incorporating reconstruction residuals

An alternative method to the above decides for an event activation, not by thresholding the elements of $\mathbf{H}'$, but by measuring KL-divergence between $\mathbf{V}$ and $\mathbf{\Lambda}$, when only the atoms of the event in question and of background noise are used in reconstruction (see Section 5.1 for details on background noise modeling). More specifically, the total reconstruction error of sparse-CNMF over a time-segment, seg, under consideration, is $D(\mathbf{V}_{seg} || \mathbf{\Lambda}_{seg})$, whereas reconstruction error on basis of only the i-th event and noise is $D(\mathbf{V}_{seg} || \mathbf{\Lambda}_{seg}^{(i,bg)})$, where,

$$\mathbf{\Lambda}_{seg}^{(i,bg)} = \sum_{t=0}^{T-1} \mathbf{W}_t^{(i,bg)} \cdot \overset{t\rightarrow}{\mathbf{H}_{seg}^{(i,bg)}} \quad , \tag{5}$$

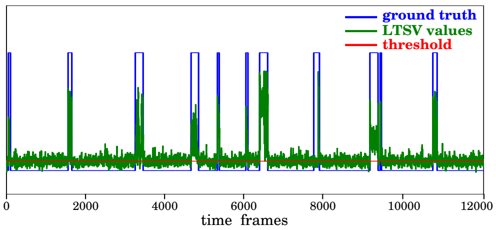

Figure 1: An example of applying the long-term signal variability (LTSV) measure to background noise detection (see Section 5.1). Ground truth peaks correspond to acoustic events.

with $\mathbf{H}_{seg}^{(i,bg)}$ denoting the part of $\mathbf{H}$ that contains only rows corresponding to atoms of the i-th class or background noise and columns that correspond to the time frames of seg. Similarly, in the above, $\mathbf{\Lambda}_{seg}$ and $\mathbf{V}_{seg}$ contain the columns of (1) and of the data matrix, respectively, within the segment under consideration.

We define the "residual ratio" of the i-th event as the ratio between the residual on basis of (5) to the total one, using (1), namely

$$\mathcal{E}(i\,,n) = \frac{D\,(\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg}^{(i,bg)})}{D\,(\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg})} \quad , \quad \text{for all} \quad n \in seg \quad . \tag{6}$$

In computing (6), non-overlapping segments of 1 sec. in duration are used. Small residual ratio values 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). Activations in $\mathbf{H}'$ with large magnitude are also often related with large percentage of reconstruction, but this is not always the case. From the minimization of (2), large magnitude activations may occur for a given event and a given time frame, but with a small corresponding reconstruction contribution.

In our first approach using activations only, the event detection criterion is the activation matrix $\mathbf{H}$ element magnitudes. In the residuals-based approach, instead, the criterion is the accuracy of reconstruction using only atoms and activations of a particular event. In our final system, submitted to the Challenge, we combine both. Thus, the i-th event is considered active at time frame n, if

$$H'(i\,,n) > \theta_H \quad \text{and} \quad \mathcal{E}(i\,,n) < \theta_{\mathcal{E}} \quad . \tag{7}$$

Thresholds θ_H and $\theta_{\mathcal{E}}$ are chosen as explained in Section 5.2.

5. SYSTEM IMPLEMENTATION DETAILS

5.1. Background noise modeling

In addition to modeling the acoustic events by incorporating representative atoms in the dictionary, background noise modeling is necessary for robust AED. With the presence of background noise atoms in the dictionary, false alarm event activations are avoided in areas that events are not present. Also, more reliable reconstruction is possible in active areas, assuming additive noise.

In our approach, and following work in [6], we extract the background noise atoms from the observed data during decoding (on-the-fly). The advantage of this scheme is the adaptation of the background dictionary to slightly different conditions, possibly existing each time. However, instead of assuming background noise present at the beginning and end of the observed data, as in [6], we attempt to extract background atoms from various areas of the signal, by employing the long-term signal variability (LTSV) measure, described in [13]. This measure has been successfully used in voice activity detection, and it is based on the fact that background noise usually exhibits smaller variability through time in its spectrum.

In our system, a frame is considered as noise if its LTSV value is lower than a fixed threshold, θ_L. As before, the shift-invariant dictionary reduction method is applied to areas that noise is detected to help provide background noise atoms. An example of the LTSV based approach is shown in Figure 1, where LTSV values for a Challenge corpus signal are depicted, together with ground-truth locations of acoustic events. As it can be seen, LTSV values and the chosen θ_L ensure that acoustic event time frames are avoided.

5.2. Features, system parameters, and post-processing

We now provide some additional details of our implemented system. Concerning audio feature extraction, we have experimented with various feature sets that satisfy non-negativity and approximate linearity: Mel-filterbank energies, Gammatone-filterbank energies, DFT spectrogram, and the variable Q-Transform (VQT). The first three are computed using 30 msec long frames with a 10 msec shift, whereas VQT is obtained from the baseline system of [8]. Our final submitted system uses 150-dimensional Mel-filterbank energy features (M= 150).

Regarding dictionary building, atoms of 200 msec ($T = 17$ frames) in duration are used, and for the CNMF-framework, parameter λ in (2) is set to 0.7. Further, approximately 200 atoms per event class are used ($R_i \approx 200$), with $R \approx 2.4$k total atoms (including background noise modeling).

Concerning the various thresholds employed, θ_H in (7) is computed as a percentage (15%) of the maximum value of matrix $\mathbf{H}'$ elements. Threshold $\theta_{\mathcal{E}}$ in (7) is computed as a percentage (106%) of the minimum of $\mathcal{E}(i\,,n)$ for a given segment. Such values are optimized on available development data (see Section 6.1).

Finally, as a post-processing stage in the detection system, one-dimensional dilation is performed on each row of matrix $\mathbf{H}'$, 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. At the end, $T - 1$ frames after each detected activation are also considered as active.

6. EXPERIMENTS

6.1. Database

We perform experiments on the DCASE'16 Challenge database designed for Task 2 – "Sound event detection in synthetic audio" [8]. The corpus contains recordings of eleven office-related acoustic events (see also Figure 2), consisting of three parts: The training set with 20 isolated recordings of each event; a development set with 18 two-minute long recordings of synthetic mixtures of audio events and noise at various SNRs and event overlap conditions ("density" and "polyphony"); and a test set of similar structure to the development set (54 recordings), only used in the Challenge evaluation, with its ground-truth publicly unavailable at the moment.

6.2. Experimental setup

In this paper, we report experiments on both the development and test sets (the latter as only provided by the Challenge organizers). Specifically, for the development set, due to its particularity of containing the same event instances as the training set, we use two different setups, described next.

Table 1: Performance of baseline and proposed systems on 3 sets.

system	setup #1		setup #2		test	
	Fscore	ER	Fscore	ER	Fscore	ER
NMF-baseline	0.42	0.79	0.32	0.87	0.37	0.89
activations-only	0.83	0.30	0.43	0.79	–	–
activations&residuals	0.84	**0.29**	0.55	**0.63**	0.56	**0.68**

Table 2: Performance of different feature sets and dictionary sizes.

features	feat. dim.	dict. size	setup #1		setup #2	
			Fscore	ER	Fscore	ER
VQT	545	200	0.79	0.37	0.29	0.88
Gamma	150	200	0.82	0.33	0.35	0.86
Mel	150	200	0.83	**0.30**	0.43	**0.79**
Mel	150	100	0.81	0.36	0.42	0.85
Mel	100	100	0.83	0.30	0.42	0.82
DFT	545	100	0.78	0.42	0.41	0.83

Table 3: Performance of different dictionary building methods.

dictionary building method	setup #1		setup #2	
	Fscore	ER	Fscore	ER
sparse-CNMF	0.64	0.60	0.29	0.89
shift-invariant reduction	0.83	**0.30**	0.42	**0.82**

- Setup #1: This is identical to the default setup of Task 2. One dictionary is built using all isolated training data, and then AED is performed on all 18 development set recordings.
- Setup #2: Here, to allow testing on unseen event instances, we perform a 18-leave-one-out experiment. In total, 18 dictionaries are built, each tested on a single development set recording, by using each time all available training set instances, except those contained in the particular development set recording.

6.3. Metrics

We report results employing the adopted Challenge metrics [8], namely frame-based Fscore and frame-based total error rate (ER). The latter is defined as $\text{ER} = (I + D + S)/N$, where I denotes acoustic event insertions, D deletions, S substitutions, and N the total number of ground-truth events at a given frame. ER is computed in frames of 1 sec. in length.

6.4. Results

In Table 1, the results using the Challenge-provided NMF baseline, our submitted system, and a variant of it are compared for the different experimental setups considered. Regarding the NMF-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, 60 bins, 10 msec step). A post-processing stage applies median filtering to the output and allows up to five concurrent events [8].

Both our systems, depicted in Table 1, perform dictionary creation employing the shift-invariant reduction approach, and their details are provided in Section 5.2. It is obvious that both outperform the baseline in all setups. In particular, our submitted system ("activations & residuals") achieves 63.3%, 27.6%, and 23.6% relative reduction in ER over the baseline for setup #1, #2, and the test set, respectively. It seems that the extraction of more atoms per class (almost ten-fold over the baseline), combined with the incorporation of temporal structure under the CNMF-framework, lead to major improvements.

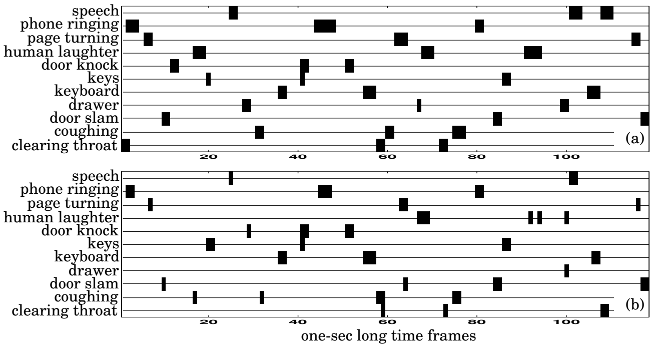

Figure 2: AED on the "dev_1_ebr_6_nec_3_poly_0.wav" Challenge recording: (a) ground-truth; (b) output of our submitted system. Acoustic event labels are also shown.

Comparing our two detection approaches, we can observe that the system using the combination of activations and reconstruction residuals (submitted to the Challenge) achieves a 20% ER relative reduction in setup #2, compared to the system using activations only. This highlights the complementarity of the two methods. The improvement is mainly due to the elimination of false activations, exhibiting large peaks in $\mathbf{H}'$ but also having a large residual ratio.

In Table 2, we show experimentation regarding different audio feature sets, together with variations in their dimensionality and dictionary size (number of atoms per class is depicted). We can observe that Mel-filterbank energies achieve the best performance among the different sets considered. It thus seems that they are more appropriate for the set of acoustic events considered in the Challenge. Also from the Mel feature results (150-dimensional), we can observe that increasing dictionary size leads to slight improvements.

A comparison of the different dictionary building methods is shown in Table 3, using the same detection system in both cases (a 100-dimensional Mel-filterbank, activations-only system, with 100 atoms per class). Clearly, the shift-invariant dictionary size reduction approach outperforms conventional CNMF-based dictionary building. This provides evidence that accurate representation of event atoms (instead of approximate) is beneficial to detection, as long as we have a way to select appropriate atoms.

Finally, in Figure 2, the output of our system is shown against ground-truth for a particular audio recording of the development set.

7. CONCLUSIONS

We presented a sparse-CNMF based system for overlapping audio event detection, employing an efficient dictionary building method and a novel detection approach. Attention was also given to background noise modeling and on experimentation with different possible feature sets for the CNMF framework. Results obtained on Task 2 of the DCASE'16 Challenge were promising, significantly outperforming the NMF-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 overlapping acoustic event detection.

8. 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] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, “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.

[5] 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.

[6] 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.

[7] 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.

[8] Detection and Classification of Acoustic Scenes and Events 2016. [Online]. Available:
http://www.cs.tut.fi/sgn/arg/dcase2016/

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.