Multi-Channel Non-Negative Matrix Factorization for Overlapped Acoustic Event Detection

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

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

pangian@cs.ntua.gr , gpotam@ieee.org , maragos@cs.ntua.gr

***Abstract*—In this paper, we propose two multi-channel extensions of non-negative matrix factorization (NMF) for acoustic event detection. The first method performs decision fusion on the activation matrices produced from independent single-channel sparse-NMF solutions. The second method is a novel extension of single-channel NMF, incorporating in its objective function a multi-channel reconstruction error and a multi-channel class sparsity term on the activation matrices produced. This class sparsity constraint is used to guarantee that the NMF solutions at a given time will contain only a few classes activated across all channels. This indirectly forces the channels to seek solutions on which they agree, thus increasing robustness. We evaluate the proposed methods on a multi-channel database of overlapping acoustic events and various background noises collected inside a smart office space. Both proposed methods outperform the single-channel baseline, with the second approach achieving a 15.4% relative error reduction in terms of F-score.**

***Index Terms*—Acoustic event detection, multi-channel fusion, non-negative matrix factorization**

I. Introduction

Acoustic event detection (AED) constitutes a significant part of computational auditory scene analysis, with the purpose of automatically detecting and identifying meaningful sound events present in an audio recording. Among others, popular applications of AED include smart home environments, multimedia indexing and retrieval [1], monitoring for healthcare [2], and security and surveillance systems [3].

Several methods have been developed in recent years for both the isolated and the more challenging overlapped AED scenarios. Hence, one can find in the literature AED systems employing hidden Markov models (HMMs) [4], probabilistic component analysis models [5], the generalized Hough-transform [6], deep neural networks [7]–[9], and non-negative matrix factorization (NMF) [10], [11], among others.

All aforementioned approaches have been primarily applied to single-channel AED. However, whenever available, exploiting information from multiple channels can be valuable. In [12] various channel fusion methods were proposed within an HMM based framework, while in [13] bag-of-words based features from different channels were used to train a global random forest classifier. Regarding neural network based methods, in [14] multi-channel exploitation was performed either by feeding the network with inputs from multiple channels or by extracting multi-channel spatial features. In NMF related approaches, multi-channel extensions have also been considered, but mostly targeting blind source separation [15]–[19].

In this paper we propose two multi-channel extensions of NMF suitable for overlapping AED. NMF-based methods, due to their natural ability to detect multiple events occurring simultaneously, have demonstrated robust performance in related tasks [20]. In this work, the single-channel baseline, upon which we build our methods, is a sparse-NMF based approach performing detection at frame-level. Our first method combines the different microphones at decision level by summing their activation matrices to obtain an average confidence for the activation of each class. Our second method considers the optimization of a novel objective function containing a multi-channel KL-divergence reconstruction term and a multi-channel class sparsity term. At each time frame, this class sparsity term forces the NMF solutions to contain only a small number of activated classes in total across all microphones. In this way, the updates of the activation matrix for each microphone at each iteration are informed by the activations from the other microphones too, and this leads to robust solutions in which most of the microphones should agree. For our experiments we use the publicly available ATHENA database [21], which contains real multi-channel recordings from a smart-office environment including sixteen acoustic events and five types of background noise. The results confirm the superiority of the proposed multi-channel approaches over the single-channel baseline.

The remainder of the paper is organized as follows: Section II presents both the single-channel baseline and the two proposed multi-channel approaches; Section III describes the database and experimental framework employed and reports our results; and, finally, Section IV concludes the paper.

II. Methods

A. Single-Channel Sparse-NMF

The main idea behind the application of sparse-NMF for AED is the linear decomposition of acoustic events into

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

ISBN 978-90-827970-1-5 © EURASIP 2018

spectral atoms. Given the representation of events with non-negative and approximately linear features (e.g., spectrogram, filterbank energies), overlapping events can be decomposed into atoms of individual events.

NMF seeks to determine a linear non-negative approximate factorization of the observed feature matrix $\mathbf{V} \in \mathbb{R}_{\geq 0}^{P \times N}$, by the product $\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$, where $\mathbf{W} \in \mathbb{R}_{\geq 0}^{P \times R}$ denotes the non-negative dictionary matrix, and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{R \times N}$ represents the non-negative activation matrix. Here P denotes the feature dimensionality, N the number of time frames, and R the total number of event atoms in the dictionary matrix. Further, sparse-NMF adds a sparsity constraint to the solution, usually requiring sparse activations in matrix $\mathbf{H}$ [22].

For the m^{th} channel, given the observed matrix $\mathbf{V}_m$ and the dictionary matrix $\mathbf{W}_m$ containing atoms for all acoustic events, sparse-NMF derives the activation matrix $\mathbf{H}_m$ by minimizing the objective function:

$$J_m = D(\mathbf{V}_m | \mathbf{W}_m \mathbf{H}_m) + \lambda \|\mathbf{H}_m\|_1 \tag{1}$$

When employing the generalized KL-divergence error cost function D, the solution to (1) can be obtained by means of the iterative update:

$$\mathbf{H}_m \leftarrow \mathbf{H}_m \odot \left\{\mathbf{W}_m^T (\mathbf{V}_m \oslash (\mathbf{W}_m \mathbf{H}_m))\right\} \oslash \left\{\mathbf{W}_m^T \mathbf{1}_V + \lambda \mathbf{1}_H\right\}$$

where $\odot$ and $\oslash$ denote element-wise matrix multiplication and division, and $\mathbf{1}_V$ and $\mathbf{1}_H$ are matrices with all elements equal to 1 and dimensions equal to $\mathbf{V}_m$ and $\mathbf{H}_m$ respectively. $\mathbf{H}_m$ is initialized with random positive values, and for its computation we apply 100 iterations.

After obtaining matrix $\mathbf{H}_m$, for each time frame, the activations for each class are summed across all their atoms resulting in a new matrix $\mathbf{H}'_m \in \mathbb{R}_{\geq 0}^{C \times N}$, where C denotes the total number of event classes. Finally, detection is performed by thresholding, i.e., class c is considered active at time frame n, if $\mathbf{H}'_m(c, n) > \theta_H$, where θ_H is a suitably chosen threshold.

Regarding the creation of dictionary matrix $\mathbf{W}_m$, we use the "exemplar" based method: Using extracted isolated training instances from each event, we create the class-specific sub-dictionaries $\mathbf{W}_m^{(c)} \in \mathbb{R}_{\geq 0}^{P \times R_c}$, for $c = 1, ..., C$, by clustering the available isolated instances with the K-means algorithm (R_c centroids are selected). The total dictionary $\mathbf{W}_m$ is then created by concatenating the C sub-dictionaries, i.e., $\mathbf{W}_m = \left[\mathbf{W}_m^{(1)}, ..., \mathbf{W}_m^{(C)}\right] \in \mathbb{R}_{\geq 0}^{P \times R}$.

B. Sum of Channel Activations

In NMF based methods, activations produced for each class are directly related to the confidence about its existence. In this multi-channel approach we combine the different channels at the decision level, expecting more reliable results, compared to that based on a single channel alone.

At first, each channel m acts independently from the others, performing single-channel sparse-NMF by using its own observation matrix $\mathbf{V}_m$ and dictionary matrix $\mathbf{W}_m$ and outputs its activation matrix $\mathbf{H}_m$. Then, the activations from all channels are averaged to obtain the final activation matrix $\mathbf{H}_f$:

$$\mathbf{H}_f = \frac{1}{M} \sum_{m=1}^{M} \mathbf{H}_m, \tag{2}$$

where M is the total number of channels considered. Finally, summing of activations per class and thresholding follow, as in the single-channel case.

C. Multi-Channel NMF with Class Sparsity

In this approach we extend the objective function of single-channel NMF in a multi-channel fashion. Towards this end, we first transform the reconstruction error term to contain the sum of KL-divergence errors from all channels. In this case, in each reconstruction term, each channel uses a global dictionary matrix $\mathbf{W}$ that is built similarly to the single-channel case approach discussed at the end of Section II.A, but with the modification that atoms are sampled for each class from all training data across all channels. Further, we add a multi-channel class sparsity constraint as a second term. This constraint is used to regularize the NMF solutions so that, at each time frame, only a few classes are activated across all channels. As a consequence, the channels are forced to act in a collaborative way and find solutions to which they agree.

The multi-channel objective function J is defined as:

$$J = \sum_{m=1}^{M} D(\mathbf{V}_m | \mathbf{W}\mathbf{H}_m) + \lambda \sum_{n=1}^{N} \Omega(h_{1,n}, ..., h_{M,n}), \tag{3}$$

where $\mathbf{H}_m = [h_{m,1}, ..., h_{m,N}]$ and $h_{m,n} = [{h_{m,n}^{(1)}}^T, ..., {h_{m,n}^{(C)}}^T]^T$, i.e., $h_{m,n}$ is the n^{th} column of the activation matrix $\mathbf{H}_m$. The class-sparsity function Ω is defined as:

$$\Omega(h_{1,n}, ..., h_{M,n}) = \sum_{c=1}^{C} \log(\epsilon + \sum_{m=1}^{M} \|h_{m,n}^{(c)}\|_1)$$

where $h_{m,n}^{(c)}$ denotes the part of the activation column that is related with the event class c. This function can be viewed as a multi-channel extension of the term used in [23], [24] to imply group sparsity. In our case groups are the event classes considered.

By majorizing the second term of (3), we obtain the following updates for activation matrices $\mathbf{H}_m$, for all $m \in \{1, ..., M\}$, $c \in \{1, ..., C\}$, $n \in \{1, ..., N\}$:

$$\mathbf{H}_m \leftarrow \mathbf{H}_m \odot \left\{\mathbf{W}^T (\mathbf{V}_m \oslash (\mathbf{W}\mathbf{H}_m))\right\} \tag{4}$$

$$h_{m,n}^{(c)} \leftarrow h_{m,n}^{(c)} \oslash \left\{(\mathbf{W}^{(c)})^T \vec{1}_v + \lambda \vec{1}_{h_c} / (\epsilon + \sum_{m'=1}^{M} \|h_{m',n}^{(c)}\|_1)\right\} \tag{5}$$

where ϵ is a small positive constant and the column vectors $\vec{1}_v$ and $\vec{1}_{h_c}$ have all their elements equal to 1 and dimensions $P \times 1$ and $R_c \times 1$ respectively. In our experiments, $\mathbf{H}_m$ is

ISBN 978-90-827970-1-5 © EURASIP 2018

initialized with random positive values, and the updates (4), (5) are applied iteratively for 100 iterations.

From (5) it can be seen that, at each iteration, the update for each channel is also affected by the activations of the other channels. In particular, when the total activation across all channels (as computed at the previous iteration) is low for a given class, the activations of the m^{th} channel at the current update are suppressed for that class. In (3), parameter λ tunes the size of the impact of this class-sparsity constraint: high values of λ will lead to solutions with only a few different classes activated at each frame.

After obtaining the M different activation matrices for all channels, we compute the final activation matrix $\mathbf{H}_f$ as in the previous method, using (2). Finally, we should note that the dictionary matrix $\mathbf{W}$ that is used in updates (4) and (5) of $\mathbf{H}_m$ has its columns (atoms) normalized so that their elements sum to 1.

It is worth mentioning that in our work we employ the multiplicative updates approach for solving the NMF task, mainly because of their widespread usage in related works and also due to their high reproducibility. Alternative efficient algorithms for solving the NMF task have also been proposed and applied successfully in the literature [25]–[27].

III. Experiments

A. Database

We perform our experiments on the ATHENA multi-modal database [21], captured in a smart office environment. In total, the dataset contains 240 one-minute long sessions of real recordings divided into a training and test set. This database is suitable for multi-channel overlapped AED, as it contains speech plus fifteen acoustic events captured from multiple microphones (20 in total) installed on the ceiling and walls of the smart space (see also Fig. 1). The acoustic events are categorized according to their average duration to long events ("walking steps", "cellphone ring", "keyboard", "glass fill", "coffee spoon", "Skype call", "cough", "paper work", "window open/close") and short events ("mouse click", "keys", "knock", "chair moving", "switch on/off", "door open/close"). To better approximate a realistic scenario, five different types of acoustic backgrounds are also considered in the various sessions (ambient noise, fan, radio music, vacuum cleaner, silence). Highly overlapped scenarios (40% of speech overlaps with other events) and adverse noise conditions make this dataset challenging for overlapped AED. The ATHENA database is publicly available [1].

B. System Implementation Details

Next, we provide details about the various parameters of the systems described. Regarding audio feature extraction, we employ 100 Mel-filterbank energies computed in windows of 30 msec duration and with a 10 msec shift. Concerning the number R_c of atoms selected per class in the dictionary

[1]http://cvsp.cs.ntua.gr/research/athenadb

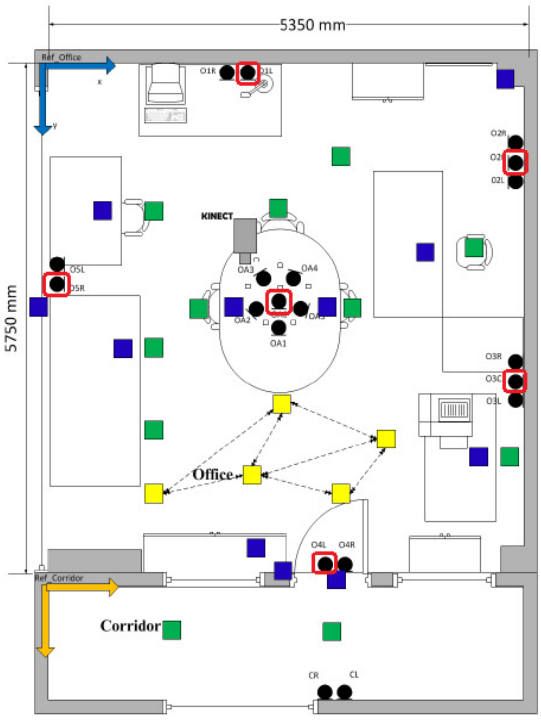

Fig. 1. Floor plan of the smart office used in the ATHENA database recordings. Microphones (black), speaker (green and yellow) and event (blue) positions are depicted. Six microphones were used in our experiments, marked with a red square (from [21]).

and the sparsity parameter λ, we experimented with various combinations: $R_c \in \{20, 40, 60, 80, 100, 120, 150\}$ and $\lambda \in \{0.5, 1, 2, 4, 8, 16, 32\}$. Also for better background modeling, we extract and store in the dictionary R_c atoms for each type of background considered.

As a post-processing stage for the detection system, after thresholding the activations with θ_H, for each class, we unify active segments that occur with time distance less than t_u sec and delete active segments with duration shorter than t_d sec. All parameters were optimized on the development set (see Section III.C).

Finally, for our multi-channel approaches we employ the six microphones that are highlighted with red marks in Fig. 1. The purpose of our selection was to uniformly sample the acoustic space.

C. Experimental Setup

In our experiments we have considered three types of acoustic backgrounds, namely ambient noise, fan, and silence, which are more common in real-life scenarios. These backgrounds cover roughly 1 hour of recordings in the training set and 1 hour in the test set. From the corresponding part of the training set, we select isolated instances of events and use them for dictionary building. We also divide the test set into development and evaluation sets, of 30 min duration each. The optimization of all system parameters was performed on the development set. The metrics used for evaluation and

ISBN 978-90-827970-1-5 © EURASIP 2018

comparison of our methods are frame based Recall, Precision, and F-score.

D. Results

The three methods are evaluated and compared in the 30 min long recordings of the evaluation set. As a baseline for our experiments we consider the average single-channel F-score, computed as the mean of the F-scores of the different single-channel NMF systems (6 in total). This corresponds to the expected performance we would get if we chose randomly a microphone in the smart space. As an alternative baseline, we also show the results of the oracle single-channel, i.e. the best-performing channel for the given evaluation set (central microphone of the ceiling in our case). In Fig. 2 the results in terms of Recall, Precision, and F-score are depicted for the baseline and the two multi-channel approaches. First, we can observe that both multi-channel approaches outperform the single-channel baseline, achieving 6.80% and 15.44% relative error reduction in terms of F-score (the sum-of-activations and multi-channel NMF methods, respectively). Further, both multi-channel methods show significant improvements over the oracle single-channel result. Also, multi-channel NMF with class sparsity performs better than the sum-of-activations method, achieving 9.27% relative error reduction. Finally, the multi-channel NMF approach shows the best results in all metrics, performing also at a slightly more balanced point between Recall and Precision than the sum-of-activations method.

We can also observe that, in general, AED performance is relatively low, indicating the challenging nature of the database. Such can be primarily attributed to the highly overlapped conditions, the adverse background noise, and the large variety of event classes considered.

Finally in Fig. 3, we can observe the effect of class sparsity parameter λ on the solutions for the activation matrices. In particular we show the activations of the final activation matrix $\mathbf{H}_f$, averaged in time, for a given time interval where two acoustic events overlap ("speech" and "cellphone ring"). We can see that, when increasing the class sparsity parameter, the solutions become more concentrated on the atoms of the given events. When λ becomes lower, atoms from more classes become activated, leading to false alarms in the detection. In the given example, when λ=16 only one false alarm occurs for event "cough", while for λ=2 false alarms also occur for events "Skype call", "window open/close", and "mouse click".

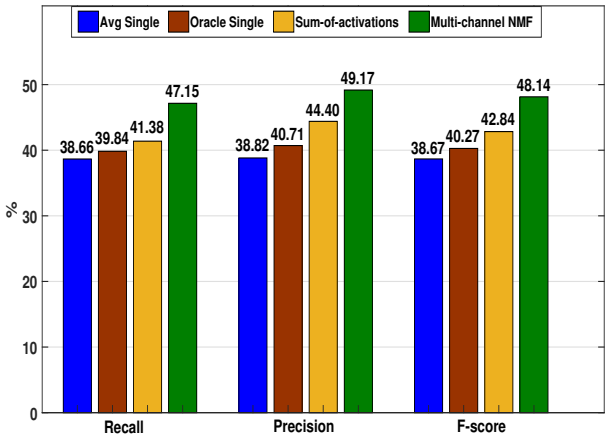

Fig. 2. AED results on the evaluation set of the ATHENA database, depicted in terms of Recall, Precision and F-score for the three different approaches of Section II.

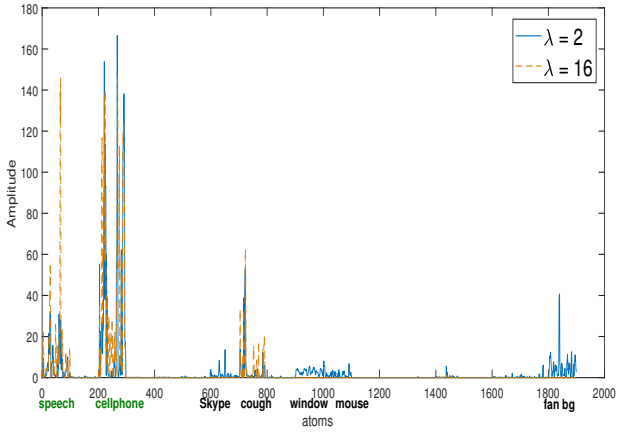

Fig. 3. Activations across atoms in activation matrix $\mathbf{H}_f$ of the multi-channel NMF method. Activations are averaged for a given time interval of 2 sec in duration and shown for two different values of sparsity parameter λ. Events with green colors overlap in the ground truth annotation of this interval.

IV. Conclusions

In this paper, we proposed two multi-channel NMF approaches for overlapped AED. The first method combines at decision level the independent sparse-NMF outputs from different channels. The second method considers the optimization of a novel multi-channel NMF objective function including a class sparsity term. Such term introduces robustness, as it forces the channels to activate only a few classes that they agree on. Both proposed multi-channel methods outperformed the single-channel baseline, with the second achieving satisfactory improvements.

In future work, we plan to investigate the performance of our methods over single-channel alternatives on additional databases, suitable for overlapped AED. We will also investigate how the number and positions of the microphones selected affect the performance of our methods. Dictionary building using multiplicative updates occurring with our sparsity term will also be tested in addition to the "exemplar" based method considered here. Finally, it will be interesting to incorporate our novel multi-channel NMF approaches into other single-channel NMF based frameworks that have been proven robust for AED, such as convolutional NMF [28].

ISBN 978-90-827970-1-5 © EURASIP 2018

REFERENCES

[1] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[2] D. Hollosi, J. Schröder, S. Goetze, and J.-E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," in *Proc. 3rd International Symposium in Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, 2010, pp. 1–5.

[3] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6460–6464.

[4] 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.

[5] E. Benetos, G. Lafay, M. Lagrange, and M. 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.

[6] 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.

[7] S. Hershey, S. Chaudhury, D. P. W. Ellis, J. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, R. A. Saurous, B. Seybold, M. Slaney, and R. Weiss, "CNN architectures for large-scale audio classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

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

[9] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *arXiv preprint arXiv:1604.07160*, 2016.

[10] 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," in *Proc. of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2016, pp. 45–49.

[11] 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.

[12] 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.

[13] J. Kurby, R. Grzeszick, A. Plinge, and G. A. Fink., "Bag-of-features acoustic event detection for sensor networks," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 55–59.

[14] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," *arXiv preprint arXiv:1706.02293*, 2017.

[15] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2018.

[16] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[17] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 261–264.

[18] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.

[19] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1942–1945.

[20] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[21] A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos, and P. Maragos, "ATHENA: A Greek multi-sensory database for home automation control," in *Proc. of 15th Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 1608–1612, [Online; http://cvsp.cs.ntua.gr/research/athenadb].

[22] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2006, pp. 2614–2617.

[23] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 293–297, 2015.

[24] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[25] N. Gillis and R. Luce, "Robust near-separable nonnegative matrix factorization using linear optimization," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1249–1280, 2014.

[26] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: An optimal gradient method for nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2882–2898, 2012.

[27] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Advances in Neural Information Processing Systems*, 2012, pp. 1214–1222.

[28] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

ISBN 978-90-827970-1-5 © EURASIP 2018