ONLINE NON-NEGATIVE TENSOR DECONVOLUTION FOR SOURCE DETECTION IN 3DTV AUDIO

Yuki Mitsufuji

Sony Corporation Tokyo, Japan

Marco Liuni[1]*, Alex Baker*[1]*, Axel Roebel*[1]

IRCAM-CNRS-UPMC UMR 9912, 75004 Paris, France

ABSTRACT

The following article describes research on source detection in multi channel (3DTV) audio streams. The problem is extremely complex due to the fact that multiple layers can be present in scenes (background music, ambience, commentator). In this work a new algorithm is developed that exploits the information from the different audio channels to detect, and possibly localize and separate independent audio sources. An algorithm based on online Non-negative Tensor Deconvolution is realized, to deal with sound sources with time dependent positions in the channel matrix. The evaluation is made on 3DTV 5.1 film soundtracks and on synthetic mixes of 3DTV 5.1 audio with target sounds from a sound effects database: a significant improvement of the detection performance is shown, compared with other decomposition techniques.

Index Terms— Dictionary training, nonnegative tensor deconvolution, source separation, event detection, 3DTV audio

1. INTRODUCTION

In the field of under-determined source separation, Non-negative Matrix Factorization (NMF, [1]) is a consolidated approach, and numerous related techniques have been devised for a variety of related problems [2, 3]. In source detection problems, some target sound events of a given class need to be identified in a complex mix. A number of source detection methods based on NMF have been proposed so far. Weninger et al. incorporate NMF-based sound event detection into a speech recognition framework to remove non-linguistic events from speech recordings [4]. There are several other sound event detection systems based on NMF, customized for realistic situations in multi source environments [5, 6]. In this paper, we discuss source detection and tracking in multichannel audio streams, and propose an extension of NMF dedicated to overcoming existing limitations of the NMF model. Our proposed extension focuses on three different aspects: the use of multichannel signals, the modeling of temporal-spectral patterns and the detection of sound objects in multichannel audio scenes. For the first aspect, Fitzgerald and others have proposed a multichannel extension of NMF called Non-negative Tensor Factorization (NTF, [7, 8]), which allows spatial positions of sources to be estimated, and has found several applications [9, 10, 11]. When dealing with detection problems in multi-channel signals, NTF presents a significant advantage over NMF after down-mixing, because it fully exploits the lower masking of targets within the individual channels. An extension to enhance detection results of NMF is to incorporate modeling of temporal-spectral pattern as a basis, as done with Non-negative Matrix Deconvolution (NMD, [12, 13, 14]). This approach has been extended to tensors in [15], with a different formalism, leading to Non-negative Tensor Deconvolution (NTD). In this paper, NTD is formulated with the formalism in [12], as it provides a direct and more efficient strategy of implementation.

This paper contains three main novelties:

1. A new approach to learn a dictionary for sound event detection that allows constraining the dictionary such that each of the elements in the dictionary is representative for a subset of the target events. The approach proposed is similar to vector quantization (VQ) of the space of properly segmented target events with the difference that for VQ amplitude scaling of events would lead to new patterns which is not desired in a NMF application.
2. A modification of the NTD model [15] that introduces the possibility of changing the coefficient in the spatial weights matrix over time, according to the so called *online approach* (see [16, 17] for the NMF case). This extension allows us to represent sound events that move within the channel matrix over time, and significantly extends the possible applications of NTD to multi channel audio as the audio sources in general can be placed in different channel positions at different times or even move between audio channels.
3. A new application of NTD based detection of real world audio events [14] in the context of multi channel audio from 3DTV media.

This paper is organized as follows: Section 2 briefly introduces NTD representation of multichannel spectrograms. Section 3 describes the system framework of the NTD-based online approach. Section 4 shows evaluation results in terms of F-measure, and lastly, Section 5 presents conclusions and remarks.

2. NONNEGATIVE TENSOR DECONVOLUTION

Let $\mathrm{V} = \{v_{ckl}\}$ be a nonnegative tensor representing a multichannel spectrogram, $\mathrm{V} \in \mathbb{R}^{+C \cdot K \cdot L}$, where C, K, L are three integers, respectively the number of channels, fftsize and number of frames of the spectrogram. In this work, a vector of a dictionary matrix in a NTF is indicated as *component*, while a matrix of a dictionary tensor in a NTD is indicated as *pattern*. Let P (number of patterns) and T (number of frames for each pattern) be two integers, and consider the nonnegative matrices $\mathrm{Q} \in \mathbb{R}^{+C \cdot P}$, $\mathrm{H} \in \mathbb{R}^{+L \cdot P}$ and $\mathrm{W}^t \in \mathbb{R}^{+K \cdot P}$, where $t \in [1, T]$ with $1 \leq T \leq L$. The NTD problem is to find a tensor $\hat{\mathrm{V}} = \{\hat{v}_{ckl}\}$ approximating V, whose coefficients are

$$\hat{v}_{ckl} = \sum_{p=1}^{P} q_{cp} \cdot \sum_{t=1}^{T} w_{kp}^{t} h_{lp}^{t \rightharpoonup} \ , \tag{1}$$

[1] This research has been partly funded by the 3DTVS project in the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287674.

where $\mathrm{H}^{t\rightharpoonup} = \{h_{lp}^{t\rightharpoonup}\}$ is defined as follows,

$$h_{lp}^{t\rightharpoonup} = \begin{cases} 0 & \text{if } 1 \leq l \leq t \\ h_{l-t,p} & \text{if } t < l \,. \end{cases}$$

The tensor $\hat{\mathrm{V}}$ should minimize the following cost function,

$$J(\mathrm{V}||\hat{\mathrm{V}}) = \sum_{c,k,l} d_\beta(v_{ckl}|\hat{v}_{ckl}) \,, \tag{2}$$

where d_β is the β-divergence [18, 19]. Equation (2) can be expanded as follows,

$$\begin{aligned} J(\mathrm{V}||\hat{\mathrm{V}}) \quad = \quad & \sum_{c,k,l} \frac{v_{ckl}^\beta}{\beta(\beta-1)} + \frac{(\sum_p q_{cp} \sum_t w_{kp}^t h_{lp}^{t\rightharpoonup})^\beta}{\beta} + \\ & - v_{ckl} \frac{(\sum_p q_{cp} \sum_t w_{kp}^t h_{lp}^{t\rightharpoonup})^{\beta-1}}{\beta-1} \,, \end{aligned} \tag{3}$$

and differentiated with respect to the coefficients of the matrices and tensor composing the approximation,

$$\begin{aligned} \nabla_Q J(\mathrm{V}||\hat{\mathrm{V}}) = & \langle \hat{\mathrm{V}}^{\circledast(\beta-1)}, \mathrm{D} \rangle_{\{2,3\};\{1,2\}} + \\ & - \langle \mathrm{V} \circledast \hat{\mathrm{V}}^{\circledast(\beta-2)}, \mathrm{D} \rangle_{\{2,3\};\{1,2\}} \,, \end{aligned} \tag{4}$$

where $\mathrm{D} \in \mathbb{R}^{+K \cdot L \cdot P}$ and $\mathrm{D}(:,:,p) = \sum_t \mathrm{W}_{:p}^t \circ \mathrm{H}_{:p}^{t\rightharpoonup}$. The $\circ$ symbol denotes the outer product, while $\langle \cdot, \cdot \rangle_{\{\cdot\};\{\cdot\}}$ is the contracted product [20]; $\circledast$ is the Hadamard product, powers of matrices indicated with $\circledast(\cdot)$ are element-wise, and the symbol $\mathrm{M}_{:p}$ selects the p-th column of the matrix M. By setting the appropriate step in a gradient descent, the following multiplicative update rule can be deduced for the matrix Q,

$$\mathrm{Q} \leftarrow \mathrm{Q} \circledast \frac{\langle \mathrm{V} \circledast \hat{\mathrm{V}}^{\circledast(\beta-2)}, \mathrm{D} \rangle_{\{2,3\};\{1,2\}}}{\langle \hat{\mathrm{V}}^{\circledast(\beta-1)}, \mathrm{D} \rangle_{\{2,3\};\{1,2\}}} \,. \tag{5}$$

In a similar way,

$$\begin{aligned} \nabla_{W^t} J(\mathrm{V}||\hat{\mathrm{V}}) = & \langle \hat{\mathrm{V}}^{\circledast(\beta-1)}, \mathrm{E} \rangle_{\{1,3\};\{1,2\}} + \\ & - \langle \mathrm{V} \cdot \hat{\mathrm{V}}^{\circledast(\beta-2)}, \mathrm{E} \rangle_{\{1,3\};\{1,2\}} \,, \end{aligned} \tag{6}$$

where $\mathrm{E} \in \mathbb{R}^{+C \cdot L \cdot P}$ and $\mathrm{E}(:,:,p) = \mathrm{Q}_{:p} \circ \mathrm{H}_{:p}^{t\rightharpoonup}$. The following multiplicative update rule can be deduced for each matrix W^t,

$$\mathrm{W}^t \leftarrow \mathrm{W}^t \cdot \frac{\langle \mathrm{V} \cdot \hat{\mathrm{V}}^{\circledast(\beta-2)}, \mathrm{E} \rangle_{\{1,3\};\{1,2\}}}{\langle \hat{\mathrm{V}}^{\circledast(\beta-1)}, \mathrm{E} \rangle_{\{1,3\};\{1,2\}}} \,. \tag{7}$$

The differentiation with respect of the coefficients in H is different, because of the action of the shift operator:

$$\frac{\partial J(\mathrm{V}||\hat{\mathrm{V}})}{\partial h_{lp}} = \sum_{ck} \sum_t q_{cp} \cdot w_{kp}^t (\hat{v}_{ck,l+t}^{\beta-1} - v_{ck,l+t} \cdot \hat{v}_{ck,l+t}^{\beta-2}) \,. \tag{8}$$

The compact form is therefore given by:

$$\begin{aligned} \nabla_H J(\mathrm{V}||\hat{\mathrm{V}}) = & \sum_t \langle (\hat{\mathrm{V}}^{\circledast(\beta-1)})^{t\leftharpoondown}, \mathrm{F}^t \rangle_{\{1,2\};\{1,2\}} + \\ & - \sum_t \langle (\mathrm{V} \circledast \hat{\mathrm{V}}^{\circledast(\beta-2)})^{t\leftharpoondown}, \mathrm{F}^t \rangle_{\{1,2\};\{1,2\}} \,, \end{aligned} \tag{9}$$

where $\mathrm{F}^t \in \mathbb{R}^{+C \cdot K \cdot P}$ and $\mathrm{F}^t(:,:,p) = \mathrm{Q}_{:p} \circ \mathrm{W}_{:p}^t$. The following update rule can finally be deduced for H,

$$\mathrm{H} \leftarrow \mathrm{H} \circledast \frac{\sum_t \langle (\mathrm{V} \circledast \hat{\mathrm{V}}^{\circledast(\beta-2)})^{t\leftharpoondown}, \mathrm{F}^t \rangle_{\{1,2\};\{1,2\}}}{\sum_t \langle (\hat{\mathrm{V}}^{\circledast(\beta-1)})^{t\leftharpoondown}, \mathrm{F}^t \rangle_{\{1,2\};\{1,2\}}} \,. \tag{10}$$

3. SYSTEM FRAMEWORK

A training/detection framework based on NTD was implemented, that is composed of three parts:

1. training a dictionary of sound patterns from a database of single-channel audio events, belonging to the target class to be detected (see section 3.1);
2. decomposing the audio scene to be analyzed, using a dictionary that includes the trained patterns and some random patterns (see section 3.2);
3. deducing a list of markers from the analysis of the obtained decomposition, which are supposed to be the start and end points of the target events within the audio scene (see section 3.3).

3.1. Dictionary Training

The patterns that are used to represent and detect the target sounds are a critical element of the detection algorithm. The dictionary of target patterns will contain P_{tar} patterns and will be denoted as $\mathrm{W}_{\mathrm{tar}}$. An ideal selection of patterns in $\mathrm{W}_{\mathrm{tar}}$ would be activated only by the target sound events and not by any other background sounds. A common strategy to find appropriate patterns or components is to find a dictionary that achieves a good representation of all target event sounds with maximally sparse (minimum L0-norm) activation. The minimum L0-norm that allows representing a collection of target events is achieved when only a single pattern or component is active at each moment.

A simple approach would be to use the target event training database as patterns in $\mathrm{W}_{\mathrm{tar}}$. In this case a perfect construction with minimum L0-norm can be achieved. The dictionary $\mathrm{W}_{\mathrm{tar}}$ may however be very large if many target event examples are used. Note that vector quantization cannot be used in the present context because intensity changes in target sound events would trigger the unnecessary construction of additional patterns. Different approaches have been proposed to build smaller target dictionaries that allow sparse representation. [14] use a sparse NMD decomposition to build the target dictionary. In [21, 6] a rank one NMF decomposition is applied to all the target events in the target event training database and from the result the target dictionary is constructed by either retaining all basis [6] or by means of calculating an average basis per target class [21].

Here we propose a new algorithm for constructing the target dictionary. The target event samples used for training are concatenated in a mono audio file, and the output dictionary $\mathrm{W}_{\mathrm{tar}}$ is computed by means of two subsequent NMDs which aim to reinforce the sparsity of the activation matrix H. The NMDs use the same cost function that is used later for the event detection. For impulsive sounds (like gunshots) each of the target events is cut to have length T starting shortly before the signal maximum amplitude measured in the time domain signal. For gunshots the frame directly before the signal maximum is used as start frame.

The first NMD is initialized such that the matrix H is non zero only in time positions forming a regular time grid that is given by the length of the dictionary patterns. This initialization ensures that no time overlap of active patterns can occur. The target dictionary is initialized with $\hat{P}_{\mathrm{tar}}$ random patterns. After convergence all but the strongest activations obtained from the first NMD at each position of the regular time are set to zero and a second NMD is computed. All patterns that are active at least once are used to construct $\mathrm{W}_{\mathrm{tar}}$. The resulting target dictionary contains then $P_{\mathrm{tar}} \leq \hat{P}_{\mathrm{tar}}$ patterns that allow a good and maximally sparse representation of the target event

sound training database. While $\hat{P}_{\mathrm{tar}}$ has to be set in advance only the necessary patterns will finally be active such that the value P_{tar} is determined adaptively as a function of the target event training sound database.

It is interesting to compare the performance of the target dictionary $\mathrm{W}_{\mathrm{tar}}$ obtained with the proposed procedure and the complete algorithm that is presented here, with the performance obtained using NMF with L1-norm sparsity constraints. The performance (F-measure) of the best dictionary obtained with this classical procedure to construct a target dictionary is about 5-10% below the performance obtained with the proposed dictionary training method. One of the problems with the sparsity constraints based on the L1-norm is the fact that the L1-norm does not allow to control sparsity as exactly as is done with the proposed algorithm.

3.2. Decomposition: the online approach

The decomposition stage is handled with NTD in a semi-supervised manner: our online approach is needed to have time-dependent information about the position of the target sources in the multi-channel mix. It consists of dividing the audio scene into consecutive segments of size L_{seg}, for each of which a new NTD is computed, obtaining a sequence of factorizations composed of the tensors $\mathrm{W}[i]$ and the matrices $\mathrm{H}[i]$ and $\mathrm{Q}[i]$, where i is the segment index. At each initialization of the iterative update, i.e. for each new segment, the activation matrix $\mathrm{H}[i]$ and the direction matrix $\mathrm{Q}[i]$ are randomly initialized. This way, the directions and activations computed for different segments are unrelated. In particular, this choice allows the direction to change freely, as different target events can, in principle, be panned anywhere in the multi-channel mix, without affecting each other. As for the tensor $\mathrm{W}[i]$ containing the dictionary, the initialization is obtained in order to account for the trained components and also to have a coherent description of the background scene when moving between the decompositions of adjacent audio segments.

The trained dictionary $\mathrm{W}_{\mathrm{tar}}$ is included within the initial dictionary $\mathrm{W}[i]$ for the online NTD, for every i. Each $\mathrm{W}[i]$ has P total components. The remaining $P - P_{\mathrm{tar}}$ components are randomly initialized at each new decomposition. The aim here is to obtain a decomposition where the activations of the target components are separate from the ones of the rest of the scene. The non-target elements of the dictionary are reinitialized at every new segment: this is motivated by the fact that the sound scene in a film can change rapidly, so that no particular continuity is required for the representation of non-target events. The information coming from different segments is later merged for the definition of a global activation matrix, described in section 3.3. The components in the target dictionary $\mathrm{W}_{\mathrm{tar}}$ are not updated during the iterative online NTD decompositions, and they should interpret all of the energy of the target events. The other components are updated according to the rules in section 2. The number P of total patterns has to be large enough to provide an exhaustive description of the audio scene, avoiding the activation of the target dictionary for non-target events.

3.3. Source Detection

For each of the P patterns, by combining the direction and activation matrices for each segment of the online NTD, a matrix $\widetilde{\mathrm{H}}_p \in \mathbb{R}^{+C \cdot L}$ is obtained, that provides information about the activation of the selected pattern in each channel over the entire scene; such matrices are defined by combining the following sub-matrices $\widetilde{\mathrm{H}}_p[i] \in \mathbb{R}^{+C \cdot L_{\mathrm{seg}}}$,

$$\widetilde{\mathrm{H}}_p[i]_{c,l} = \mathrm{Q}[i]_{c,p}\, \mathrm{H}[i]_{l,p}\ , \tag{11}$$

$$\widetilde{\mathrm{H}}_p = \left[\widetilde{\mathrm{H}}_p[1] \mid \widetilde{\mathrm{H}}_p[2] \mid \widetilde{\mathrm{H}}_p[3] \mid ... \mid \widetilde{\mathrm{H}}_p[I]\right]\ ,$$

where I denotes the total number of segments. As detailed in the following section, this is the only information used to decide if a certain segment of the analyzed scene contains a target event.

The average of the per-channel activation of the target patterns is computed, and a decision system with a double threshold determines the start and end points of detected events. The first threshold determines if the activation value is high enough to assign that analysis frame to a target event in a certain channel, and to set a starting point. After the first threshold has been reached, a second threshold is used to set the end of the target event, when the activation falls below it. Depending on the target sound, a filter to fix the minimum and maximum duration of each detection is applied.

4. EVALUATION

There are many advantages in conceiving a framework for multichannel nonnegative decomposition, instead of applying single-channel techniques on down-mixes of the multichannel sound file, or repeatedly on the separated channels. Compared to a single-channel downmix of the audio scene, multi-channel techniques can exploit the enhanced readability of sound objects in the separated channels, as well as the information coming from the inter-channel relations. Moreover, the computational cost of a NTD is significantly less than the cost of NMD on all channels independently. The error calculation is about the same for both cases, but independent NMD for 5 channels has 5 times more parameters (activations and basis) to treat. The mixing matrix Q in NTD does not add significantly to the costs. Additionally, to be able to fuse information from different channels, additional computation has to be performed with NMD. About accuracy, the NTD model fits the audio scene (one sound source projected into all channels) and is therefore conceptually easier than independent NMD on individual channels. With sources in multiple channels, NTD can provide fusion between channels, but NMD cannot (or less easily).

For the evaluation of the proposed framework, gunshot and car engine target sounds were considered. The algorithm described in section 3.3 has been implemented in a framework where both NTF or NTD can be used. Gunshots can be more easily detected than cars with single components, i.e. individual spectrogram frames. As such, this case is evaluated here for both the NTF and NTD decompositions, to show the advantage gained from the use of spectral patterns instead of components. The Itakura-Saito divergence is used for the iterative minimization of the error function in the NTF and NTD decompositions.

For the evaluation, existing 3D TV film audio was used, containing sufficient target sounds, together with artificial mixes of 3D TV film audio with added target sounds. Several scenes have been analyzed from the following films: *Le Fabuleux Destin d'Amélie Poulain* (2001, directed by Jean-Pierre Jeunet); *The Dark Knight* (2008, directed by Christopher Nolan); *Drive Angry* (2011, directed by Patrick Lussier). They shall be referred to as *Amelie*, *Batman* and *Drive*, respectively. Table 1 contains the scores of the system in each situation described in the following sections, in terms of accuracy, precision, recal and F-measure.

4.1. Artificial and original mixes: trained and untrained events

Several artificial mixes have been realized, adding gunshots or car engine samples, from sound databases, to the soundtracks of *Amelie* and *Batman*. Some mixes use the 22 gunshots and 33 car engines

Mix	Duration	Target	% of target	Type	Method	Thresholds	Accuracy	Recall	Precision	F-Measure
Amelie	2'44"	gunshots	9.4	trained	NTF	opt	48.76	70.23	61.45	65.55
					NTD	opt	71.27	79.76	87.01	83.22
			10.26	untrained	NTF	fixed	49.57	71.08	62.1	66.29
						opt	50	72.28	61.85	66.66
					NTD	fixed	63.82	72.28	84.5	77.92
						opt	64.77	68.67	91.93	78.62
		cars	15.87	trained	NTD	opt	93.03	79.58	77.2	78.37
			18.51	untrained	NTD	fixed	91.50	89.61	71.61	79.6
						opt	92	86.29	74.54	79.99
Batman	4'52"	gunshots	10.34	untrained	NTF	fixed	30.29	36.05	65.43	46.49
						opt	40.18	61.22	53.89	57.32
					NTD	fixed	57.34	81.2	66.12	72.89
						opt	61.2	75.16	76.71	75.93
		cars	18.66	untrained	NTD	fixed	65.58	84.73	32	46.46
						opt	84.64	49.27	56.97	52.84
Drive	1'55"	gunshots	7.3	untrained	NTF	fixed	3.61	100	3.61	6.98
						opt	22.87	57.14	27.58	37.2
					NTD	fixed	12.14	92.85	12.26	21.66
						opt	23.07	64.28	26.47	37.5
		cars	25.32	untrained	NTD	fixed	62.4	85.17	38.93	53.43
						opt	83.14	65.21	67.25	66.21

Table 1. Results of the detection based on the NTF and NTD online approaches (see section 4).

that have been used in the training stage: they are indicated with *trained* and are used to test the system's ability to retrieve the trained objects when mixed in a given background. Other mixes, indicated with *untrained*, use target sounds from different databases, and are meant to test the performance of the system with generic targets, as well as the dependence of the algorithm on the initial dictionary.

The audio from the scene in *Drive* contains both gunshots and cars, and so is used as it is, to test the performance of the system in a real-world use case. These cases are still indicated as *untrained*.

The number of elements in the dictionaries is $P_{\mathrm{tar}} = 10$ with $P = 50$ for the NTF case with gunshots, while it varies depending on the target events in the NTD case: $P_{\mathrm{tar}} = 10$, $P = 20$ and $T = 20$ for gunshots, $P_{\mathrm{tar}} = 27$, $P = 40$ and $T = 20$ for car engines.

4.2. Adaptive thresholds training and score evaluation

The *Amelie* scene and the corresponding *trained* mixes are also used to train the double-threshold for the decision routine in the detection stage. The score is obtained with brute force optimization of the activation and deactivation thresholds based on the F-measure, and is indicated as *opt*. These thresholds are two multiplying factors applied to the average of the $\widetilde{\mathrm{H}}_p$ matrices corresponding to the target elements in the dictionary: they vary adaptively, depending on the presence of the specific target in a mix and its prominence.

For the other mixes, the thresholds obtained in the corresponding *trained* case with the *Amelie* scene are first used, to check the robustness of the procedure: the scores obtained this way are indicated with *fixed*. Then, to evaluate the dependence of the thresholds on the changing background, the same brute force optimization is applied, and the corresponding score is also indicated with *opt*.

The use of NTD improves the F-measure obtained with the NTF in all the cases. For gunshots, both systems show a good level of independence from the trained database, as shown by the scores of the *Amelie* scene in the untrained tests, some of which are even higher than those of the trained case.

Changing the background, the scores decrease variably depending on the target. For gunshots, the principal confusion is introduced by percussive events in the soundtrack, which are prevalent in *Drive*, where the score is extremely low. This problem can be solved with a refinement of the training stage, by using discriminant techniques [22] to exclude certain event classes from the potential targets. For cars, the problem comes from the broad variety of the class, while the training stage is performed over a database limited to car engines: a more exhaustive dataset would improve the scores, that are already globally satisfying. A further reason to explain the low scores in *Drive* is that the audio scene of this film is extremely dense, and hard to resume in a standard annotation based on precise occurrences (start/end point) of sound objects: the evaluation presented in table 1 is based on a one-frame precision compared to the annotation. This is done for clarity purposes, but a standard content-based research would not need such a high accuracy, and could therefore be based on the analysis of activations over a longer term, consistently improving the scores.

The differences between the *opt* and *fixed* scores show that the adaptive thresholds training is often satisfying, considering that a single scene is used as reference: these differences can be drastically reduced if thresholds are trained on a larger variety of scenes, with different backgrounds.

5. CONCLUSIONS AND FUTURE WORK

The developed framework based on online NTD shows good results for the task of source detection in multi channel (3DTV) audio streams. In relation to the exposed work, the ongoing works are focused on three main aspects: the inclusion of discriminant techniques [22] in the training stage to improve the selectivity of the initial dictionary. Then, the choice of larger sound banks, allowing an exhaustive representation of the different targets. Finally, an optimization of the parameters in each stage is investigated, together with the reduction of the decomposition to selected portions of the spectrum, in order to reduce the still significant computational cost of the framework.

6. REFERENCES

[1] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[2] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2003 IEEE Workshop on*, oct. 2003.

[3] Tuomas Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *International Computer Music Conference*, 2003.

[4] F. Weninger, B. Schuller, M. Wollmer, and G. Rigoll, "Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, may 2011, pp. 5840 –5843.

[5] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Proc. of Intern. Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, 2011, pp. 36–40.

[6] Arnaud Dessein, Arshia Cont, and Guillaume Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, pp. 341–371. Springer, Berlin, Germany, 2012.

[7] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *ICML*. 2005, vol. 119, pp. 792–799, ACM.

[8] D. FitzGerald, M. Cranitch, and E. Coyle, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. of the Irish Signals and Systems Conf. (ISCC) 2005*, 2005.

[9] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[10] A. Ozerov, C. Fevotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, may 2011, pp. 257 –260.

[11] Joonas Nikunen, Tuomas Virtanen, and Miikka Vilermo, "Multichannel audio upmixing based on non-negative tensor factorization representation," in *WASPAA*, 2011, pp. 33–36.

[12] Paris Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, CarlosG. Puntonet and Alberto Prieto, Eds., vol. 3195 of *Lecture Notes in Computer Science*, pp. 494–499. Springer Berlin Heidelberg, 2004.

[13] C. Lopes and F. Perdigão, "Speech event detection by non-negative matrix deconvolution," in *Proc. of European Signal Processing Conference (EUSIPCO 2007)*, 2007.

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

[15] D. FitzGerald and M. Cranitch, "Sound source separation using shifted non-negative tensor factorisation," in *Proc. of the IEEE Conference on Audio and Speech Signal Processing (ICASSP*, 2006.

[16] Cyril Joder, Felix Weninger, Florian Eyben, David Virette, and Björn Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *LVA/ICA*, 2012, pp. 322–329.

[17] Z. Duan, G. Mysore, and P. Smaragdis, "Online plca for real-time semi-supervised source separation," *Latent Variable Analysis and Signal Separation*, pp. 34–41, 2012.

[18] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Signals and Systems Conference (ISSC 2009), IET Irish*, june 2009, pp. 1 –6.

[19] M. Nakano, H. Kameoka, J. Le Roux, Yu. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, 29 2010-sept. 1 2010, pp. 283 –288.

[20] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, Wiley. com, 2009.

[21] Jouni Paulus and Tuomas Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proceedings of the 13th European Signal Processing Conference*, 2005, p. 4.

[22] I. Kotsia, S. Zafeiriou, and I. Pitas, "A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems," *Information Forensics and Security, IEEE Transactions on*, vol. 2, no. 3, pp. 588–595, 2007.