Sparse representations for modeling environmental acoustic scenes, application to train stations soundscapes

Benjamin Cauchi, Mathieu Lagrange, Nicolas Misdariis, Arshia Cont

ACOUSTICS 2012

Sparse representations for modeling environmental acoustic scenes, application to train stations soundscapes

B. Cauchi, M. Lagrange, N. Misdariis and A. Cont

Institut de Recherche et Coordination Acoustique/Musique, 1, place Igor Stravinsky 75004 Paris
cai@idmt.fraunhofer.de

Our daily life happens in a world of dense acoustic environments. This is especially the case in train stations, where soundscapes are usually very complex. In this paper, we will investigate how Non Negative Matrix Factorization methods can be used to obtain a low rank spectrogram approximation, composed of spectral templates that can be related to some salients events like footsteps, whistles, etc... We thus assume here that the scene can be characterized by a few salient events that occur several times within the scene. We also assume that even if the acoustic realizations of those events cannot be considered in isolation, those realizations have similar spectro temporal properties. We here consider 66 recordings made in French train stations, where individual salient events have been manually annotated. We then assess the ability of the methods to extract meaningful components by comparing the activations of those components within the scene to the manually annotated ones. Experiments demonstrate that enforcing sparsity on the activations, i.e. constraining that only a few components is active at a time, has a positive effect.

1 Introduction

Detecting events of interest within auditory scenes is an interesting problem among CASA studies. Though human audition is able to detect such events within complex sound mixtures the tasks becomes quite difficult for algorithms when applied to real auditory scenes with strong background noise.

Non-negative Matrix Factorization (NMF) is an approach introduced by Lee & Seung [5] in which the data is described as the product of a set of basis and of a set of activation coefficients both being non-negative. As the data is constructed additively, NMF can provide a meaningful representation of an auditory scene. NMF and its various extensions have been proven efficient in sources separation [1] [8], real-time pitch detection in musical content [3] and supervised detection of acoustic events [2].

We argue in this paper that NMF is a viable method to detect events within sound mixtures without any prior knowledge of their content. Moreover, the sparseness constraint introduced by P. Hoyer [4] in the NMF framework seems to be a convenient criterion to discriminate between salient events and background noise as a high sparseness would imply a source significantly active during short periods of time.

We first present the NMF algorithm in section 2. In section 3, we illustrate the event detection achieved by NMF on simple artificial scenes and introduce some metrics to evaluate the achieved detection. In section 4, we apply sparse NMF to a corpus of soundscapes of train station from the perceptive study of J. Tardieu [7]. We study the influence of the sparseness constraint described by P.D. O' Grady [6] and propose a selection of the elements of dictionary using the sparseness constraint from P. Hoyer.

2 Sparse Non-negative Matrix Factorization

2.1 Method

Non-negative matrix factorization (NMF) is a low-rank approximation technique for multivariate data decomposition. Given an $n \times m$ real non-negative matrix $\mathbf{V}$ and a positive integer $r < min(n, m)$, it aims to find a factorization of $\mathbf{V}$ into an $n \times r$ real matrix $\mathbf{W}$ and an $r \times m$ real matrix $\mathbf{H}$ such that:

$$\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H} \tag{1}$$

The multivariate data to decompose is stacked into $\mathbf{V}$, whose columns represent the different observations, and whose rows represent the different variables.

NMF is an iterative process that can be used in supervised or unsupervised learning. The learning is considered supervised when the dictionary $\mathbf{W}$ is given and not updated along the iterations. In this case, it is usually built beforehand as the concatenation of spectral vectors representative of each present source. In the unsupervised case, no prior information about the content is available and $\mathbf{W}$ is randomly initialized and updated along with $\mathbf{H}$. In realistic scenarios, building the inputed $\mathbf{W}$ would require to collect relevant recordings of the desired sources and to build a new $\mathbf{W}$ for each application. In the contrary, a reliable unsupervised algorithm would not require to collect any learning data and could be more easily applied to a wider range of applications.

At each iteration, the process aims at reducing a cost function C. In this work, we use a generalized version of the Kullback Leibler divergence as our cost function:

$$C(\mathbf{V}, \mathbf{WH}) = \left\| \mathbf{V} \otimes \log \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\| \tag{2}$$

Where the multiplication $\otimes$ and the division are elementwise. The rank r of the factorization corresponds to the number of elements present in the dictionary $\mathbf{W}$.

In the case of information extraction from audio files, $\mathbf{V}$ could be the amplitude of the spectrogram and therefore, $\mathbf{W}$ would be a basis of spectral features when $\mathbf{H}$ would represent the levels of activation of each of those features along time. NMF is here used to extract emerging events relevant to the classification task. Those events are expected to be significantly present but during a limited time interval. One convenient way of representing such expectation is to add a sparseness constraint on the activation coefficients within $\mathbf{H}$.

2.2 Sparseness Constraint

The very definition of sparseness (or sparsity) is that a vector is sparse when most of its elements are null. In its application to NMF, the addition of a sparseness constraint λ permits to trade off between the fitness of the factorization and the sparseness of $\mathbf{H}$. We use the NMF implementation[1] of O'Grady sparse convolutive NMF that can be used for both the convolutive extension of the NMF algorithm or the multiplicative update used in this work [6]. The sparseness constraint results in the new cost function:

$$C(\mathbf{V}, \mathbf{WH}) = \left\| \mathbf{V} \otimes \log \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\| + \lambda \sum_{ij} \mathbf{H}_{ij} \tag{3}$$

With the norm of each of the objects within $\mathbf{W}$ fixed to unity.

[1] http://ee.ucd.ie/~pogrady/scNMF/

3 Event Detection on artificial scenes

3.1 Artificial scenes

This experiment illustrates the event detection achieved with NMF and the influence of the sparseness constraint on the achieved performances. Eight artificial scenes of 15 seconds duration have been created. Four of those scenes are composed of drum sounds, chosen because of their low non-stationarity, and are referred to by *Drums*. The four others are constructed such as to be closer from what we would expect in real-life auditory scenes (such as voice, bell ring or dog barking) and are referred to by *Realistic*.

All the sound sources come from mono files encoded at 44100 Hz. Each scene is created by the addition of four tracks containing one sound repeated several times. A binary truth vector is associated to each track and is equal to one when the source is active and to zero otherwise. In order to evaluate the robustness of the algorithm, pink noise is added to each scene with an Signal to Noise Ratio (SNR) of 0.1 dB and 10 dB, referring to the energy of the signals. The scenes with added noise are referred to by subscript indices.

3.2 Evaluation of the sources detection

The Receiver Operating Characteristic (ROC) curve is a well known tool to evaluate the performances of a two group classification task. This experiment aims to label each source as active or inactive for each sample of the input signal. The Area Under the ROC Curve (AUC) can therefore be established for each track, considering the lines of $\mathbf{H}$ as the score. When the dictionary $\mathbf{W}$ is input, learned in a supervised manner, each line of the extracted $\mathbf{H}$ is matched with the corresponding line of the binary truth. In the case of unsupervised learning, as no knowledge of the organization of $\mathbf{W}$ and $\mathbf{H}$ is available, some distinction has to be made with regard to the matching established between the time varying coefficients in $\mathbf{H}$ and the lines of the binary truth $\mathcal{T}$. Two evaluations are established for the unsupervised case:

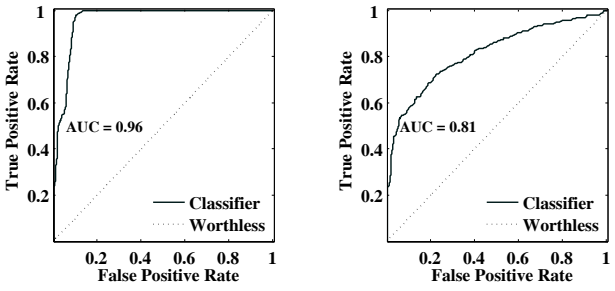

Figure 1: ROC curves of the detection of the tom in scene D1, clean (left) and with added pink noise at SNR=10dB

- AUC_{bp}: evaluates the AUC of each track for all the possible permutations of the lines of $\mathbf{H}$, and keep the best global score.
- AUC_{oh}: evaluates the distance between each row of $\mathbf{H}$ and each line of the binary truth. The AUC considered for each line of $\mathbf{H}$ is the one considering the closest line of the binary truth.
- AUC_{ohr}: evaluates the AUC similarly to AUC_{oh}. However, in this case each line of $\mathbf{H}$ can be matched only once with a line of the binary truth. The AUC of each track is computed starting with the line of the binary truth containing the highest number of active samples.

AUC_{bp} may provide artificially good results for each scenarios, making the comparison difficult. AUC_{oh} allows each line of $\mathcal{T}$ to be matched with several lines of $\mathbf{H}$. AUC_{ohr} circumvent this issue and as a consequence is used in section 4.

3.3 Description of the experiment

NMF is applied on the spectrogram of each scene computed using the short-time Fourier transform with a Hamming window of length 1024 and an overlap of 50%. The order r of the factorization is set to $r = 4$ when no noise has been added and to $r = 5$ when it has. The sparseness constraint λ has been set to 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.99 in order to evaluate its influence on the achieved source detection.

Two cases are studied here, supervised and unsupervised. In the supervised case, along with $\mathbf{V}$, a previously learned $\mathbf{W}$ is input to the NMF algorithm. Each element of this dictionary $\mathbf{W}$ has been learnt by applying the NMF algorithm with $r = 1$ and $\lambda = 0$ to the spectrogram of the audio file of each separated source. As $\mathbf{H}$, as well as $\mathbf{W}$ in the unsupervised case, are randomly initialized, ten runs have been done with each set of parameters to gain statistical significance.

3.4 Results

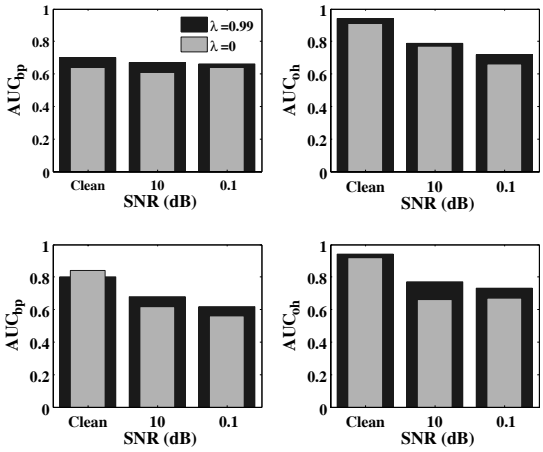

Figure 2: AUC_{bp} and AUC_{oh} achieved in the unsupervised case for *Drum* (top line) and *Realistic*, with $\lambda = 0$ and $\lambda = 0.99$

AUC_{oh} is higher than AUC_{bp} as show in Figure 2. It seems that even in the simple case of those artificial scenes and with $r = 4$, the elements to be detected are represented using several elements of $\mathbf{W}$, thus indicating that, even in a detection framework, a correct modeling shall require many more spectral vectors than sources. Indeed, as the spectral content is not stationary, it seems logical that it could not be represented by only one weighted spectral vector.

4 Event detection in complex auditory scene

4.1 Corpus and experiment

The corpus is made of scenes recorded by J. Tardieu in his study of the human perception of similarity between soundscapes of train stations [7]. It is composed of 66 audio files recorded in 6 different train stations. The recordings have been made in six types of spaces within each of those stations: platform, hall, corridor / stair, waiting room, ticket office, shop.

For each of those scenes, the list of the recognizable sound sources present during the recordings is provided in [7]. For each scene, the time interval during which the above mentioned events are present have been manually annotated. This annotation provides a binary presence indicator for each sound source used as ground truth, similarly to the experience described in 3.

4.2 Experiment

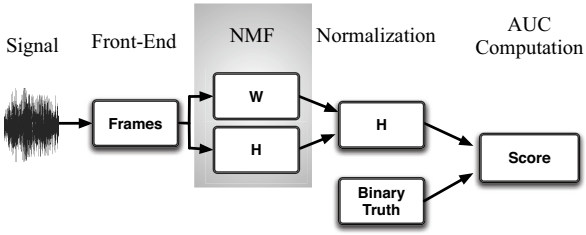

Figure 3: Overview on the process applied to each of the 66 auditory scenes

This experiment aims to evaluate the influence of sparseness on the relevance of the extracted elements of dictionary present in **W** by considering the process summarized in Figure 3.

NMF is applied on the spectrogram of each scene computed using the short-time Fourier transform with a Hamming window of length 1024 and an overlap of 50%. The order r of the factorization is set to $r = 10$, $r = 25$ and $r = 50$, as the number of actual sources is unknown. The sparseness constraint λ has been set to 0, 0.5,0.8,and 0.99 in order to evaluate its influence on the event extraction. As no knowledge of the spectral content of the auditory scenes is available, NMF is used in the unsupervised case, both **W** and **H** being randomly initialized.

The extracted **W** and **H** are normalized such that:

$$\begin{aligned} \mathbf{W}(i) &= \frac{\mathbf{W}(i)}{\sum_{f=1}^{n} \mathbf{W}(i)} \\ \mathbf{H}(i) &= \mathbf{H}(i) \times \sum_{f=1}^{n} \mathbf{W}(i) \end{aligned} \tag{4}$$

Where i is either a column of **W** or a line of **H**. AUC_{ohr} is finally computed using the extracted **H** as the score and the binary truth from the annotation as the target.

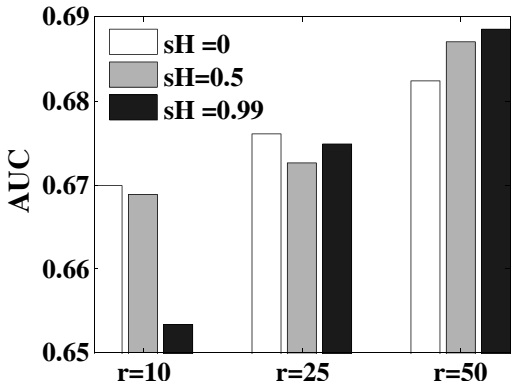

Figure 4: means of the scores achieved for all scenes and all considered events for different values of r and λ

4.3 Results

4.3.1 Sparseness Influence

Figure 4 represents the average of the AUC achieved for all of the considered events on the evaluated scenes. The best fit between the binary truth and **H** is achieved for the highest order of factorization, $r = 50$. Meanwhile, the sparseness constraint is counterproductive for a low r but slightly improves the performances when the order of factorization has been set high enough. This observation can be interpreted intuitively. As the sparseness constraint enforces the elements of **W** to be less active other time, more elements of dictionary have to be extracted in order to obtain meaningful components.

4.3.2 Resynthesis using sparseness selection

As the extracted **W** and **H** can be used to reconstruct the scene, a proper selection among the extracted elements may highlight the salient events of interest. Though, we cannot evaluate its merit within a denoising framework as no clean reference signal is available for this corpus.

As stated before, the sparseness over time λ_t proposed by Hoyer seems an interesting measure to establish such discrimination:

$$\lambda_t(\mathbf{H}(i)) = \frac{\sqrt{n}}{\sqrt{n}-1} \frac{\|\mathbf{H}(i)\|_1}{\|\mathbf{H}(i)\|_2} \qquad \forall i \in [1, r], 0 \leq \lambda_t(\mathbf{H}(i)) \leq 1 \tag{5}$$

As the salient events are present during a short interval, it can be expected that they would be represented by elements of dictionary with sparse activation. In order to illustrate that phenomenon, the scenes have been reconstructed using the reconstructed **W** and **H**, in which the 5 less sparse over time elements have been neglected, as filter gain applied to the spectrogram of the scene. The figure 5 represents the achieved reconstruction when applied to a recording from a platform of the station Avignon TGV. Between the 3rd and 4th second, the whistle is more salient while the background noise has been significantly reduced[2].

[2]The original recording and its reconstruction are available at `http://recherche.ircam.fr/equipes/analyse-synthese/lagrange/research/nmfAcoustics`

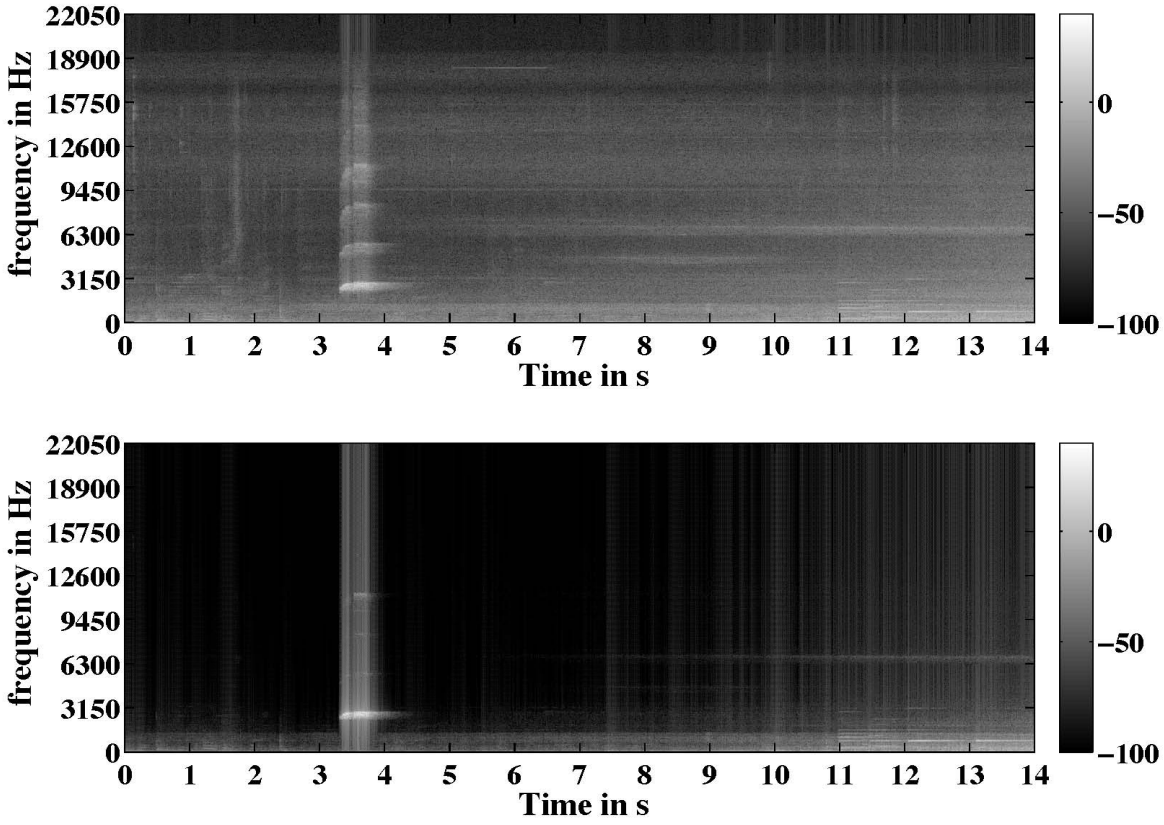

Figure 5: Spectrogram of a recording from a train station platform, with scale in dB. Between seconds 3 and 4, the whistle is more salient and the background noise is reduced from the original (top) to the reconstructed using selected elements of dictionary

5 Conclusion

Imposing sparseness within the NMF algorithm improves the detection of the interval of salient events both in the simple artificial scenes and on the real auditory scenes. In the case of the auditory scenes, however, this is the case only if the rank of the factorization is high enough. This fact can be explained by the need to fully express the complexity of the scene. Also, we have shown that using temporal sparseness as a discrimination criterion permits to reduce background noise and to increase the saliency of events of interest.

Further work would include analysis of the influence of the different parameters and particularly of the ratio between the order of the factorization and the number of selected elements for a possible reconstruction.

References

[1] A. Cichocki, R. Zdunek, and S. Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2004.

[2] 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*, pages 69 –72, oct. 2011.

[3] A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proc. 11th International Society for Music Information Retrieval Conference (ISMIR2010)*, 2010.

[4] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[5] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001.

[6] P.D. O'grady and B.A. Pearlmutter. Convolutive non-negative matrix factorisation with a sparseness constraint. In *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, pages 427–432. IEEE.

[7] J. Tardieu, P. Susini, F. Poisson, P. Lazareff, and S. Mcadams. Perceptual study of soundscapes in train stations. *Applied Acoustics*, 69(12):1224–1239, 2008.

[8] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1066–1074, 2007.