Saliency-based modeling of acoustic scenes using sparse non-negative matrix factorization

Benjamin Cauchi, Mathieu Lagrange, Nicolas Misdariis, Arshia Cont

SALIENCY-BASED MODELING OF ACOUSTIC SCENES USING SPARSE NON-NEGATIVE MATRIX FACTORIZATION

† *Benjamin Cauchi,* ‡ *Mathieu Lagrange,* ‡ *Nicolas Misdariis, and* ‡ *Arshia Cont*

† FRAUNHOFER IDMT
Hearing, Speech and Audio Technology
Marie-Curie-Str, 2,
26129 OLDENBURG - GERMANY
benjamin.cauchi@idmt.fraunhofer.de

‡ IRCAM CNRS UPMC
1, place Igor Stravinsky,
75004 PARIS - FRANCE
firstname.surname@ircam.fr

Abstract

The modelling of auditory scenes is a challenging task in Computational Auditory Scene Analysis. A method based on sparse Non-negative Matrix Factorization that can be used with no prior knowledge of the audio content to establish the similarity between scenes is proposed. The method is evaluated on a corpus of soundscapes of train stations issued from a perceptual study and results are compared with the human perception. The proposed method, by being able to focus on salient events within the scene, achieves better performances than a state-of-the-art Bag-of-Frames approach though not reaching the human performances.

1. INTRODUCTION

Relevant models of auditory scenes are needed in numerous Computational Auditory Scene Analysis (CASA) studies [1]. The input signal being mostly unconstrained in terms of content and structure, this task is a real challenge. Standard approaches, such as the Bag-of-Frames (BOF) considered in [2], aims at modeling the timber similarity between acoustic scenes and feed spectral features computed on a frame by frame basis from the composite signal to a statistical model without previous separation of the elements. Even though this scheme may be effective in some scenarios it can lack robustness when the scene is noisy and composed of only a few salient elements. Human perception is able to classify efficiently this kind of auditory scenes even in such drastic conditions. One reason for this is that human subjects identify the type of auditory space they are listening to according to some audio features (e.g. reverberation) and to salient sound sources [3]. Therefore, it seems relevant that a computational method able to extract salient events from a sound mixture would be closer to the human process both in terms of functional design and performances. For that purpose, Non-negative Matrix Factorization (NMF) seems like a relevant approach to consider. NMF has been introduced in [4]and describes data as the product of a set of basis and of a set of activation coefficients, both being non-negative, hopefully providing a meaningful and compact description of the scene. Its implementation described in [5] has been used in several applications such as supervised detection of acoustic events [6]. It has been shown to provide sparse representation of soundscapes in [7] using the same corpus as in this work. A method based on sparse NMF is proposed to model the acoustic scene with a few components that will put the focus of the model on the salient parts of the input signal. This method assumes no prior knowledge of the audio content. Both the BOF and the NMF algorithm are briefly described in section 2 and the proposed method is introduced in section 2.3. This method is then benchmarked with a corpus issued from a perceptual study [8] and shows promising improvement over the BOF method.

This work was partially funded by the European Commission under grant no. 284628 S4ECoB - Sounds for Energy Control of Buildings and the French Agence Nationale de la Recherche (ANR) under reference ANR-11-JS03-005-01

2. AUDITORY SCENES SIMILARITY

2.1. The Bag-of-Frames approach

The BOF models the time distribution of spectral features in order to establish timber similarity along time. This algorithm is shortly summarized here. The spectral envelope of each of the audio files to be classified is represented by its Mel-Frequency Cepstral Coefficientss (MFCCs). The distribution of those MFCCs over time is modeled using a Gaussian Mixture Model (GMM). GMMs estimates a probability density as the sum of $\mathcal{M}$ gaussians, where $\mathcal{M}$ is an integer parameter that has to be set. Thereof, if v is a MFCCs vector its probability density $p(v)$ can be model as:

$$p(v) = \sum_{m=1}^{\mathcal{M}} \pi_m \mathcal{N}(\mu_m, \Sigma_m) \quad (1)$$

Were $\mathcal{N}$ is a Gaussian of mean μ_m with a covariance matrix Σ. Each Gaussian of the mixture is weighted by its prior

probability π_m. Those mixture parameters need to be estimated. Typically this estimation is done using an Expectation Maximization (EM) algorithm. The similarity matrix is finally obtained by computing the distance between the GMM estimated for each file. Let a and b be two scenes with their associated probability distribution being p_a and p_b.The distance $\gamma(a, b)$ between a and b is:

$$\gamma(a,b) = \int p_a(t) \log \frac{p_b(t)}{p_a(t)} \mathrm{d}t \tag{2}$$

The BOF has been shown useful in its application to the classification auditory scenes [2] where it gave good performances on a corpus of urban soundscapes.

2.2. NMF-based proposed method

2.2.1. The NMF algorithm

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

$$\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H} \tag{3}$$

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 input $\mathbf{W}$ 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. With the multiplication $\otimes$ and the division being element-wise operations and the rank r of the factorization corresponding to the number of elements in the dictionary $\mathbf{W}$. The algorithm can be summarized as follows:

$$\begin{aligned}
&\text{Given} && \mathbf{V} \in \mathbb{R}_+^{n_f \times n_t}, r \in \mathbb{N}^* \; s.t. \; r < min(n_f, n_t) \\
&\text{apply} && \begin{cases} \mathbf{W} = \mathbf{W} \otimes \frac{\mathbf{V}}{\mathbf{WH}} \cdot \mathbf{H}^T \\ \mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}^T \cdot \frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T \cdot \mathbf{1}} \\ \text{with } \mathbf{1} \text{ a } m \times n \text{ matrix of ones} \end{cases} \\
&\text{subject to} && \mathbf{W} \in \mathbb{R}_+^{n_f \times r}, \; \mathbf{H} \in \mathbb{R}_+^{r \times n_t} \\
&\text{to minimize} && \mathcal{C}(\mathbf{V}, \mathbf{WH}) \; w.r.t. \; \mathbf{W}, \mathbf{H}
\end{aligned} \tag{4}$$

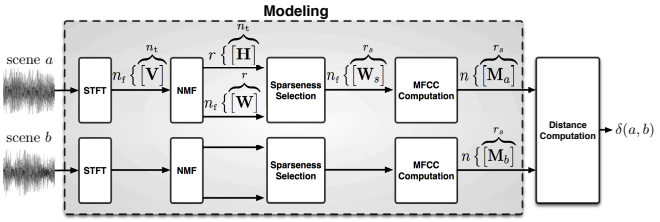

Fig. 1. Overview of the NMF based modelling. Each scene is finally represented by a matrix of only $n \times r_s$ coefficients independently of the length of the original audio.

Where $\mathcal{C}$ is the cost function that the algorithm aims to reduce. The three most common cost functions used in the context of NMF are the Frobenius norm, the Itakura-Saito divergence and the Kullback-Leibler Divergence (KLD). The cost function used in this work extends the KLD in which the right term enforce a sparseness constraint minimizing the L_1-norm of $\mathbf{H}$:

$$\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \left\| \mathbf{V} \otimes \log \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\| \tag{5}$$

$$\mathcal{C}(\mathbf{V}, \mathbf{WH}) = \mathcal{D}(\mathbf{V}, \mathbf{WH}) + \lambda \sum_{ij} \mathbf{H}(i,j) \; , \lambda \text{ a constant} \tag{6}$$

It can be noted that in Equation 6, $\lambda \sum_{ij} \mathbf{H}(i,j)$ depends only of $\mathbf{H}$. Thereof, if $\mathbf{W}$ is scaled up while $\mathbf{H}$ is scaled down, $\mathbf{WH}$ would remain unchanged while $\mathcal{C}(\mathbf{V}, \mathbf{WH})$ would decrease. To avoid the cost function to decrease while actually not obtaining a more accurate reconstruction, the elements within $\mathbf{W}$ have to be normalized to have their norm fixed to unity [5]. The iterative process can be stoped by applying a criterion on $\mathcal{C}$ or by setting up the number of iteration to be executed.

2.3. Proposed Method

The method used NMF in order to model acoustic scenes before computing pairwise distances. An overview of this modelling is presented on Figure 1 and the method can be summarized as follow. For each of the auditory scene, the spectrogram, *i.e.* the magnitude of the Short Time Fourrier Transform (STFT), is computed and constitutes a data matrix $\mathbf{V}$ of n_f rows and n_t columns. $\mathbf{V}$ is input in the NMF algorithm described in Section 2.2.1 that provides for each scene a dictionary matrix $\mathbf{W}$ composed of r spectral vectors and a matrix $\mathbf{H}$ composed of r time activation vectors. In the following, $\mathbf{W}(i)$ is the i^{th} column of the dictionary $\mathbf{W}$ while $\mathbf{H}(i)$ is the i^{th} row of the activation matrix $\mathbf{H}$. Those elements are chosen according to the sparseness sp of their respective activations along time according to the definition introduced in [9] and presented in Equation 7 where $\| \cdot \|_p$ is the p-norm.

$$\begin{aligned} sp(\mathbf{H}(i)) &= \frac{\sqrt{n} - \|\mathbf{H}(i)\|_1 \; / \; \|\mathbf{H}(i)\|_2}{\sqrt{n} - 1} \\ \forall i \in [1, r], 0 &\leq sp(\mathbf{H}(i)) \leq 1 \end{aligned} \tag{7}$$

The asumption is that the most relevant elements of $\mathbf{W}$ will be the ones for which the respective time activations will be the most sparse. The scene can then be modeled by $\mathbf{W}_s$ that consists of the r_s elements of $\mathbf{W}$ for which the respective activations within $\mathbf{H}$ are the most sparse. $\mathbf{W}_s$ is finally warped to the Mel scale on which a Discrete Cosine Transform (DCT) is applied to reduce each of its columns to n MFCCs. Before applying the DCT, the potential very small values of each $\mathbf{W}_s(i)$ are replaced by its median. This matrix of MFCC is noted $\mathbf{M}$ and is the representation used to establish pairwise distances between scenes. Let a and b are two scenes modeled by the respective dictionaries of MFCCs being $\mathbf{M}_a$ and $\mathbf{M}_b$. Computing the distance between a and b requires an operator invertible and invariant to the possible permutations between the elements of $\mathbf{M}_a$ or $\mathbf{M}_b$. Thereof, $\mathcal{E}$ being the euclidean distance between two vectors, the distance $\delta(a, b)$ between the scenes a and b is defined as:

$$\delta(a,b) = \sum_{k=1}^{r_s} \min\left\{\mathcal{E}(\mathbf{M}_a(k), \mathbf{M}_b(l))\ ,\ l \in [1, r_s]\right\} \quad (8)$$

3. CORPUS AND EXPERIMENT

3.1. Sound Corpus

The corpus used in this work is composed of 66 acoustic scenes recorded by J. Tardieu in a study of the human perception of similarity between soundscapes of train stations. The recordings and the perceptive study are detailed in [8]. The recordings have been made in six different train stations and are classified in 6 groups according to the type of space in which they have been recorded. This space typology is composed of: platform, hall, corridor / stair, waiting room, ticket office, shop and constitute the ground-truth used in the experiment. The perceptual study contains a 6 choices forced-categorization task in which the subjects have been asked to place each of the 66 audio files in one of the categories labeled according to the space typology. The subjects showed around a 40 % error rate which shows that this is a challenging corpus even for human subjects.

3.2. Experiment

Both the proposed method and the BOF approach allow to establish pairwise distances between acoustic scenes and thereof can produce a similarity matrix. In the case of the BOF, the method summarized in 2.1 has been applied using the implementation provided in [1] with the settings described in [2]. The method proposed in 2.3 is applied by varying the parameters λ and r introduced in Section 2.2.1. λ has been set to 0, 0.5, 0.8 and 0.99 while r has been set to 10, 25 and 50. The iterative process has been set to stop after 15 iterations independantly of the value of $\mathcal{C}$. The performance scores have been computed after the 5th, 10th and 15th iteration in order to observe the influence of the number of iterations. As we apply unsupervised NMF, both $\mathbf{W}$ and $\mathbf{H}$ are randomly initialized for each scene with normally distributed pseudo-random values. Thereof, in the case of the proposed method, the scores are the mean of 100 trials. When representing the extracting dictionary with MFCCs, the number of coefficients has been set to 13 as it is a common value in many applications. The influence of this MFCC step along with the one of the other parameters is discussed in Section 4. The results are given using the the 5-Precision (5-P) and the Mean Average Precision (MAP). The 5-P is the precision at rank 5 while the MAP is the average of the average precision for a set of queries.

[1] http://www.jj-aucouturier.info/projects/mir/index.htm

	BOF	NMF with r set to:			Human
		10	25	50	
5-P	0.18	0.40	0.44	0.45	0.73
MAP	0.24	0.26	0.29	0.31	0.62

Table 1. Results using NMF are the ones obtained for $\lambda = 0.99$ which gave the best scores. Best performance are achieved for $r = 50$.

4. RESULTS AND OBSERVATIONS

The Table 1 summarizes the results achieved using the BOF and the proposed NMF based method for different order of factorization r. It appears that the proposed method shows promising improvement over the BOF approach though it does remain quite inferior to the human perception. It also appears that representing the dictionary by MFCCs not only reduces the computation required to determine the distances but provides as well a slight improvement ($.02$ MAP improvment for $r = 50$).

A high constraint λ improves the performances of the algorithm but does so only if the order of factorization r has been set high enough. As shown in Figure 2, for $r = 50$ the performance increases with λ however, for a lower value of r, this constraint can be counterproductive. As the cost function $\mathcal{C}$ is guaranteed to converge it could seem that a higher number of iterations cannot hurt the performances. However, it appears that the performances improve until the 10th iteration but are poorer at the 15th one as seen in Figure 3 (left). This result suggests that an early stop around ten iterations is beneficial both in terms of computational cost and performance.

In order to examine the sparseness selection described in 2.3, the 5-precision has been computed for the 1st to rth most sparse elements. Figure 3 (right) represents the achieved score considering all of the elements and the selections of the most sparse elements that achieve the best performance for $r = 10,\ 25$ and 50. Though the selection based on the sparseness over time improves the performances its impact is less important than the of the one of the parameters λ and r.

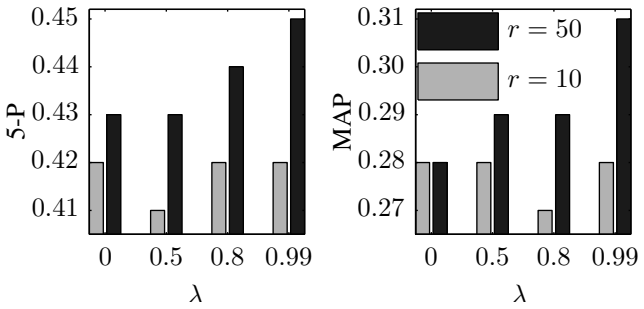

Fig. 2. The influence of constraint λ depends of the order of factorization r. A high λ improves performances if r is set high enough but decreases them if r is too low.

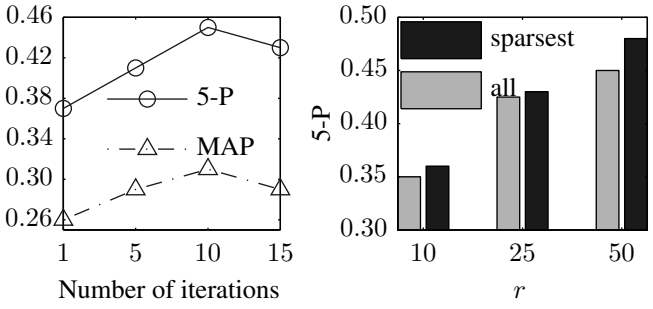

Fig. 3. Left, 5-P and MAP along the iterations. Too many iterations is counterproductive. Right, 5-Precision considering either all of the elements or the best selection of the most sparse elements.

5. CONCLUSION

The proposed NMF-based method allows to model and compare acoustic scenes without any prior knowledge about the specific characteristics of the events of interest. Experiments showed promising improvements on a challenging corpus and demonstrated that enforcing sparsity within the NMF update is useful to enhance the detection of salient events provided that the number of components is sufficiently high. It compares favorably to the BOF approach that is unable to cope with high background noise and reverberation that mask the events of interest in the spectral envelope.

Unfortunately, the performances remain far below the ones achieved by human subjects. Obviously, a strong handicap of the mentioned computational methods is the fact that they are completely unsupervised while humans can use extensive prior-knowledge about audio sources they may encounter. One way of tackling this problem is to input some abstract knowledge about the sources of interest which are expected to be present in the scene by modelling them as Time / Frequency patches as opposed to instantaneous spectral bases to better model the non stationary behavior of sources. In this respect, we believe that it would be interesting to investigate whether those priors can be described parametrically in terms of abstract shape and duration or have to be learned over some training samples in order to improve performance.

6. REFERENCES

[1] D.L. Wang and G.J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, IEEE Press, 2006.

[2] J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music.," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[3] E. Rosch, *Principles of categorization*, pp. 189–206, MIT Press, 1999.

[4] D D Lee and H S Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788791, 1999.

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

[6] 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.

[7] B. Cauchi, Lagrange, M., N. Misdariis, and A. Cont, "Sparse representations for modeling environmental acoustic scenes, application to train stations soundscapes," in *Proc. 11th French Congress of Acoustics, Nantes, France.*

[8] J. Tardieu, P. Susini, F. Poisson, P. Lazareff, and S. Mcadams, "Perceptual study of soundscapes in train stations," *Applied Acoustics*, vol. 69, no. 12, pp. 1224–1239, 2008.

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