UPMC PARIS UNIVERSITAS

PARCOURS MASTER 2
ATiAM
Parcours multi-mentions du Master Sciences et Technologies
Université Pierre et Marie Curie - Paris 6
en collaboration avec TELECOM ParisTech et l'Ircam

ircam Centre Pompidou

Master's Thesis

Non-Negative Matrix Factorization Applied to Auditory Scenes Classification

Benjamin Cauchi

August 2011

Summary

This master's thesis is dedicated to the automatic classification of auditory scene using non-negative matrix factorization. A particular attention is paid to the performances achieved by the non-negative matrix factorization in sound sources detection. Our intuition was that a good classification could be achieve if we could efficiently detect the sources within auditory scenes. It appears on short artificial examples that taking into account the non-stationarity of the spectral content of the sound sources improves the source detection. Finally, our classification method is applied to a corpus of soundscapes of train stations and the results are compared with previous classifications methods. We finally conclude that using non-negative matrix factorization significantly improves the classification.

Ce rapport de master est dédié à la classification automatique de scènes sonores utilisant la factorisation en matrices non-negatives. Une attention particulière est portée aux performances de la factorisation en matrices non-negatives dans le cadre de la détection de sources sonores. Notre intuition première a été qu'une classification performante pourrait être réalisée grâce une détection de sources efficace. Il s'est révélé sur de courts exemples artificiels que la prise en compte de la non-stationarité du contenu spectral des sources sonores améliore la détection de sources. Enfin, notre méthode de classification a été appliquée à un corpus de paysages sonores de gares et les résultats ont été comparé à d'autres méthodes de classification. Nous avons finalement conclu que l'utilisation de la factorisation en matrices non-négatives améliore significativement la classification.

Acknowledgements

I would like to thank my tutor at IRCAM, Mathieu Lagrange for having proposed this subject, his advices, teaching and patience.

I would also like to acknowledge Nicolas Misdariis, Arnaud Dessein, Arshia Cont and the Perception and Sound Design team for their help.

I would like to thank Julien Tardieu and Jean Julien Aucouturier who have been kind enough to answer my questions about their previous works.

Contents

Introduction

Non-negative matrix factorization is a technique for data decomposition and analysis. The main philosophy of this technique is to build up the observed data in an additive manner, so that cancellation is not allowed. The technique has been applied to various problems such as face recognition [12], semantic analysis of text documents [18] or audio analysis [16]. The present work has two distinct objectives. First, it aims to illustrate the source detection achieved by NMF and how the set parameters and the type of algorithm influence the performance. In a second time, making the hypothesis that the similarity between auditory scenes comes from the different sources within, it aims to evaluate the relevance of NMF as a tool for unsupervised classification of auditory scenes.

We study the source separation achieved by different standard NMF algorithms on artificial auditory scenes, which allow us to compare supervised and unsupervised learning. In our application of the NMF, we consider that the learning is unsupervised when we have no prior knowledge of the present sources and supervised when spectral content representative of each source is input as a dictionary. It appears that in both cases, standard NMF is quite efficient for a low level of background noise but achieves poor performances when the signal to noise ratio increases. We have assumed that taking into account the non stationarity of the sounds encountered in everyday life could improve the performances and therefore, we applied the extension of the convolutive NMF, proposed in [13], to the source separation in complex auditory scene. This Model is well suited to take into account the temporal evolution of the spectrum of the sources along time. It has been efficiently applied in supervised source separation of musical content [17].

In a second time, we study how NMF algorithms can be used as a unsupervised tool in order to classify complex auditory scene. In this case, the factorization does not aim to establish an accurate source separation but to extract features relevant to the classification task. We compare the performances achieved the NMF algorithms with the classification obtained via a perceptual study and with the results obtained in CASA by the bag-of-frame approach [1].

1 Presentation of the Non-negative Matrix Factorization

1.1 Non-negative matrix factorization

1.1.1 Definition

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:

$$\mathbf{V} \approx \mathbf{WH} \tag{1.1}$$

The multivariate data to decompose is stacked into $\mathbf{V}$, whose columns represent the different observations, and whose rows represent the different variables. NMF can be used in supervised or unsupervised learning. In our applications, the learning is considered as unsupervised when no prior knowledge of the sources is known, with $\mathbf{W}$ and $\mathbf{H}$ randomly initialized, and as supervised when $\mathbf{W}$ is input and that each of its columns is a representation of a source we aim to identify. Each column $\mathbf{v}_j$ of $\mathbf{V}$ can be expressed as:

$$\mathbf{v}_j \approx \mathbf{W}\mathbf{h}_j = \sum_i h_{ij}\mathbf{w}_i \tag{1.2}$$

Where $\mathbf{h}_j$ and $\mathbf{w}_i$ are respectively the $j-th$ column of $\mathbf{H}$ and the $i-th$ column of $\mathbf{W}$. The columns of $\mathbf{W}$ then form a basis and each column of $\mathbf{H}$ is the decomposition or encoding of the corresponding column of $\mathbf{V}$ into this basis. The rank r of the factorization is generally chosen such that $(n+m)r \ll nm$, so $\mathbf{WH}$ can be thought of as a compression or reduction of $\mathbf{V}$.
In the case of information extraction from audio files, $\mathbf{V}$ could be the amplitude of the spectrogram and therefore, $\mathbf{H}$ would be a basis of spectral features when $\mathbf{H}$ would represent the levels of activation of each of those features along time. The data in $\mathbf{V}$ is supposed to be non-negative, and the basis and activation coefficients are constrained to be non-negative. Therefore, cancellation is not allowed in NMF and if the input $\mathbf{V}$ is a spectrogram NMF provides a reasonable approximation of an additive representation. As auditory scenes are composed of the addition of different sound sources, NMF seems well suited to extract meaningful features in application to CASA.

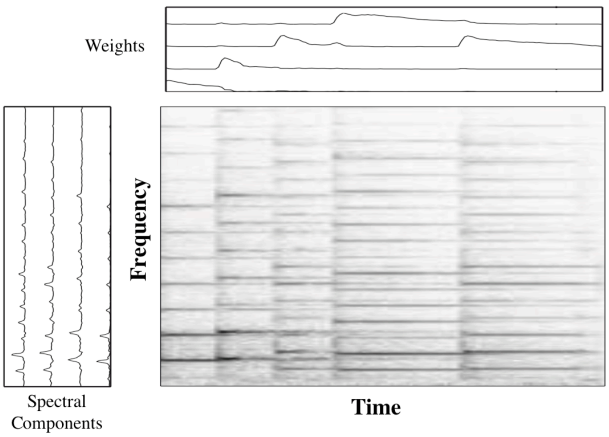

Figure 1: Illustration of standard NMF applied to audio spectrum [9]

1.1.2 Standard problem

NMF algorithms are iterative process resulting in a factorization **WH** that may be inexact, i.e. differ from **V**, so the factorization is only approximate. The aim is then to find the best factorization with respect to a given goodness-of-fit measure $\mathcal{C}$ called cost function or objective function. In the standard formulation introduced by Lee & Seung [7], the Frobenius norm is used to define the following cost function:

$$\mathcal{C}(\mathbf{W},\mathbf{H}) = \frac{1}{2}\sum_j \|\mathbf{v}_j - \mathbf{W}\mathbf{h}_j\|_2^2 = \frac{1}{2}\sum_{i,j}(v_{ij} - [\mathbf{WH}]_{ij})^2 \tag{1.3}$$

Thus, the NMF optimization problem can be expressed as:

$$\begin{array}{ll} \text{Given} & \mathbf{V} \in \mathbb{R}_+^{n\times m}, r \in \mathbb{N}^* \ s.t.\ r < min(n,m) \\ \text{minimize} & \frac{1}{2}\|\mathbf{V} - \mathbf{WH}\|_F^2 \ w.r.t.\ \mathbf{W},\mathbf{H} \\ \text{subject to} & \mathbf{W} \in \mathbb{R}_+^{n\times r},\ \mathbf{H}\ \in \mathbb{R}_+^{r\times m} \end{array} \tag{1.4}$$

The uniqueness of the solution of the equation 1.10 has to be considered up to a permutation of the lines of **H** and columns of **W**, and up to a diagonal rescaling. Even then, the solution is not unique. This is due to the fact that the cost function $\mathcal{C}$ is convex neither in **H** nor in **W**. Several definitions of $\mathcal{C}$ have been used in the literature in order to improve the exactness of the solution. In the implementation of NMF algorithms, a stop criterion $s_{\mathcal{C}}$ has to be set in order to stop the iterative process. This stop criterion can be a value of $\mathcal{C}$ for which we consider that the factorization has reached a sufficient fitness. However, as one can be unsure about the value

that $\mathcal{C}$ may reach, and that it decreases slower along the iterations, $s_{\mathcal{C}}$ may be the difference between two successive values of $\mathcal{C}$. It avoids using too much iterations for a too small improvement in the fitness of the solution.

1.2 Common Algorithms

Alternating least squares

The alternating least squares algorithms were the first to be used to solve NMF problems [11]. The idea is to update $\mathbf{W}$ and $\mathbf{H}$ in turn by minimizing $\mathcal{C}$ respectively w.r.t. $\mathbf{W}$ or $\mathbf{H}$ until convergence. For the first update, either $\mathbf{W}$ or $\mathbf{H}$ needs to be initialized.

$$\mathbf{H} \leftarrow \underset{\mathbf{H}\in\mathbb{R}_+^{r\times m}}{\operatorname{argmin}} \|\mathbf{V} - \mathbf{WH}\|_F^2 \qquad \mathbf{W} \leftarrow \underset{\mathbf{W}\in\mathbb{R}_+^{r\times m}}{\operatorname{argmin}} \|\mathbf{V} - \mathbf{WH}\|_F^2 \quad (1.5)$$

Gradient descent

The gradient descent algorithms are a particular case of *additive updates* algorithms whose principle is to give additive update rules so as to progress in a direction, called learning direction, where the cost function $\mathcal{C}$ is decreasing. In gradient descent, the learning direction is expressed using the gradient of $\mathcal{C}$. For the standard NMF problem, the following additive update rules can be deduced for the coefficients of $\mathbf{W}$ and $\mathbf{H}$:

$$h_{ij} \leftarrow h_{ij} - \mu_{ij}\frac{\partial \mathcal{C}(\mathbf{W},\mathbf{H})}{\partial h_{ij}} \qquad w_{ij} \leftarrow w_{ij} - \nu_{ij}\frac{\partial \mathcal{C}(\mathbf{W},\mathbf{H})}{\partial w_{ij}} \quad (1.6)$$

where $\mu_{ij} >= 0$ and $\nu_{ij} >= 0$ are the respective learning rates or steps of progression of h_{ij} and w_{ij}. The gradient coordinates are given by:

$$\frac{\partial \mathcal{C}(\mathbf{W},\mathbf{H})}{\partial h_{ij}} = \left[\mathbf{W}^T\mathbf{WH} - \mathbf{W}^T\mathbf{V}\right]_{ij} \qquad \frac{\partial \mathcal{C}(\mathbf{W},\mathbf{H})}{\partial w_{ij}} = \left[\mathbf{WHH}^T - \mathbf{VH}^T\right]_{ij} \quad (1.7)$$

The main problem of the gradient descent algorithms is the choice of the steps. Indeed, they should be small enough to reduce the cost function, but not too small for quick convergence.

Multiplicative updates

The multiplicative updates algorithms for NMF were introduced by Lee & Seung [6], [7] as an alternative to the additive updates algorithms such as gradient descent. The multiplicative updates are however derived from the gradient descent scheme, with judiciously chosen descent steps that lead to the following update rules:

$$h_{ij} \leftarrow h_{ij} \times \frac{\left[\mathbf{W}^T\mathbf{V}\right]_{ij}}{\left[\mathbf{W}^T\mathbf{W}\mathbf{H}\right]_{ij}} \qquad w_{ij} \leftarrow w_{ij} \times \frac{\left[\mathbf{W}\mathbf{H}^T\right]_{ij}}{\left[\mathbf{W}\mathbf{H}\mathbf{H}^T\right]_{ij}} \tag{1.8}$$

Like in gradient descent, these update rules are applied in turn until convergence. To avoid potential divisions by zero and negative values due to numerical imprecision, it is possible in practice to add a small constant ε to the numerator and denominator, or to use the non-linear operator $max(x, \varepsilon)$.

Compared to gradient descent algorithms, multiplicative updates are easy to implement and guarantee the non-violation of the non-negativity constraints if $\mathbf{W}$ and $\mathbf{H}$ are initialized with non-negative coefficients. However, despite [**?**] claims that multiplicative updates converge to a local minimum of the cost function, several authors remarked that the proof shows that the cost function is non-increasing under these updates, which is slightly different from the convergence to a local minimum [**?**]. Compared to alternating least squares algorithms, multiplicative updates are computationally more expensive and undergo slow convergence time. Finally, since a null coefficient in $\mathbf{W}$ or $\mathbf{H}$ remains null under the updates, the algorithm can get stuck into a poor local minimum.

1.3 Sparseness

The simplest definition of sparseness (or sparsity) is that a vector is sparse when most of its elements are null. In its application to NMF, improving the sparseness of the different rows of $\mathbf{H}$ helps to make the different elements of $\mathbf{W}$ specific to one source. No consensus on how sparseness should actually be defined and measured, with the result that numerous sparseness measures have been proposed. We use the sparseness introduced by Hoyer in the context of NMF [**?**]. Let X be a vector of length n :

$$sp(x) = \frac{\sqrt{n}||X||_1/||X||_2}{\sqrt{n}-1} \tag{1.9}$$

$sp(x)$ is comprised between 0 for any vector with all components equal up to the signs, and 1 for any vector with a single non-null component, interpolating smoothly between the two bounds.

1.3.1 Horizontal Sparseness

Projected gradient optimization has been used by Hoyer [5] to control sparseness in NMF. It enforces additional constraints on $\mathbf{W}$ and/or $\mathbf{H}$, more precisely to enforce $\mathbf{W}$ and/or $\mathbf{H}$ to have a desired sparseness $s_p(\mathbf{W}) = s_w$, $s_p(\mathbf{H}^T) = s_h$,with $0 \leq sw,\ sh \leq 1$ chosen by the user. The optimization problem is then:

$$\begin{array}{ll} \text{Given} & \mathbf{V} \in \mathbb{R}_+^{n \times m}, r \in \mathbb{N}^* \ s.t.\ r < min(n,m) \\ & s_w \text{ and/or } s_h \text{ s.t. } 0 \leq s_w,\ s_h \leq 1 \\ \text{minimize} & \frac{1}{2}\|\mathbf{V} - \mathbf{WH}\|_F^2 \ w.r.t.\ \mathbf{W}, \mathbf{H} \\ \text{subject to} & \mathbf{W} \in \mathbb{R}_+^{n \times r},\ \mathbf{H}\ \in \mathbb{R}_+^{r \times m}\ , \|\mathbf{w}\| = 1\ \forall j \\ & s_p(\mathbf{W}) = s_w \text{ and/or } s_p(\mathbf{H}^T) = s_h \end{array} \tag{1.10}$$

In Hoyer's work, the sparseness constraint is applied separately on each line of the matrix $\mathbf{H}$. It forces the activation coefficients to be sparse along time. It prevents the elements of the dictionary $\mathbf{W}$ to be active all the time but does not preclude them to be active at the same time. We will call this application of the sparseness constraint the horizontal sparseness and will refer to its factor by S_h.

1.3.2 Vertical Sparseness

The sparseness constraint has been as well applied in order to prevent the different elements of the dictionary to be active at the same time. A. Cont has combined it with the gradient descent update algorithm in the application to real-time multipitch observation. P. O'Grady [9] described an extension of the convolutive NMF algorithm that imposes a sparseness constraint between the different rows of $\mathbf{H}$. A. Dessein [3]has studied how to control the achieved sparseness.

1.4 Temporality and NMF

For most of the sound sources encountered in auditory scenes, as for example music notes or speech, stationarity of the spectral content only holds for one frame. However, in the standard NMF, assuming that the order r of the NMF is set to the number of sources present in the scene, the contribution of each source to the spectrum of the scene is modeled by a single vector of spectral features for which only the weight is varying along time.

In the standard NMF as described in 1.1, one way of dealing with the non-stationarity of audio would be to learn a large dictionary, containing several components per source. It could reduce the cost function and do a better job in reconstructing the nuances in the sound. However it would not produce any grouping of the dictionary components and in the case of a sound mixture we would not be able to identify which element corresponds to which source.

The Convolutive NMF [9] [13] and the Non-negative Hidden Markov Model [4] are two methods which has been developed in order to take into account the non-stationarity and improve the the performances achieved in sources separation.

1.4.1 Convolutive NMF

The Convolutive NMF extends the NMF method by using a 3 dimensional tensor $\mathbf{W}$ as dictionary. For each object i, the corresponding element $\mathbf{W}_i$ of dictionary is a a sequence of successive spectral features vectors and the corresponding extracted activation pattern $\mathbf{H}_i$ represents the starting points at which $\mathbf{W}_i$ is superposed to recreate the contribution of the i^{th} source . The generative model of 1.1 is extended to the convolutive case:

$$[!h]\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}} \tag{1.11}$$

where$\mathbf{V} \in \mathbb{R}^{>0,M\times N}$ is the input to be decomposed,$\mathbf{W}_t \in \mathbb{R}^{>0,M\times R}$and $\mathbf{H} \in \mathbb{R}^{>0,R\times N}$ are its two factors, and T is the length of each spectrum sequence. The i$^{\text{th}}$ column of $\mathbf{W}_t$ describes the spectrum of the i$^{\text{th}}$ object t time steps after the object has begun. The $\overset{i\rightarrow}{(\cdot)}$ denotes a column shift operator that moves its argument i places to the right, as each column is shifted off to the right the leftmost columns are zero filled. Conversely, the $\overset{\leftarrow i}{(\cdot)}$ operator shifts columns off to the left, with zero filling on the right.

Using the previously presented framework for NMF, the new cost function for the convolutive model is:

$$D(\mathbf{V}\|\mathbf{\Lambda}) = \left\| \mathbf{V} \otimes \log \frac{\mathbf{V}}{\mathbf{\Lambda}} - \mathbf{V} + \mathbf{\Lambda} \right\| \tag{1.12}$$

where $\mathbf{\Lambda}$ is the approximation to $\mathbf{V}$ and is defined as:

$$\mathbf{\Lambda} = \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}} \tag{1.13}$$

s

This new cost function can be viewed as a set of T conventional NMF operations that are summed to produce the final result. Consequently, as opposed to updating two matrices ($\mathbf{W}$ and $\mathbf{H}$) as in conventional NMF, $T+1$ matrices require an update, including all $\mathbf{W}_t$ and $\mathbf{H}$. The resultant convolutive NMF update equations are:

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}_t^T \cdot \overset{\leftarrow t}{[\frac{\mathbf{V}}{\mathbf{\Lambda}}]}}{\mathbf{W}_t^T \cdot 1} \qquad \mathbf{W}_t = \mathbf{W}_t \otimes \frac{\frac{\mathbf{V}}{\mathbf{\Lambda}} \cdot \overset{t\rightarrow T}{\mathbf{H}}}{1 \cdot \overset{t\rightarrow T}{\mathbf{H}}} \tag{1.14}$$

At each iteration $\mathbf{H}$ and all $\mathbf{W}_t$ are updated, where $\mathbf{H}$ is updated to the average result of its updates for all $\mathbf{W}_t$. Therefore, we can note that if $T = 1$ it is equivalent to the standard multiplicative updates.

The convolutive extension of the NMF provides an easy to implement method of dealing with the temporal variation of the spectral content. However, it still has some important limitations. The time-length of the objects to be extracted has to be input and as this time length is fixed, the convolutive NMF does not allow time stretching or other kinds of modulation of the dictionary spectral content.

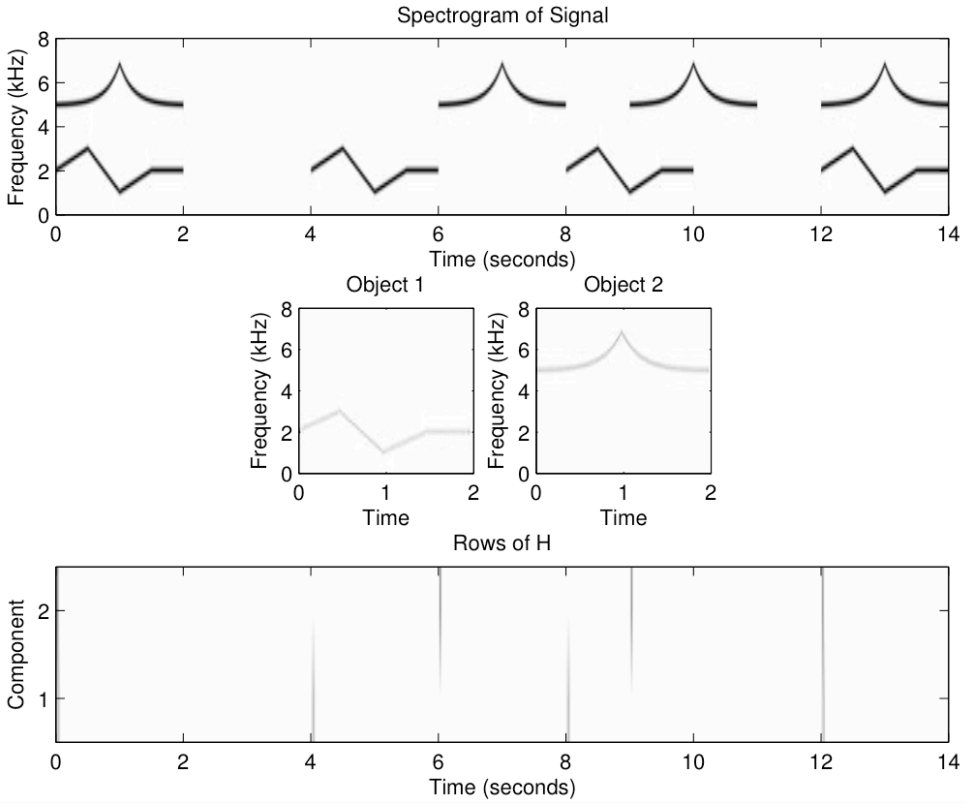

Figure 2: Illustration of Convolutive NMF applied on a two sources sound mixture [9]

1.4.2 Non-negative Factorial Hidden Markov Model

N-HMM to Model Single Sources [8]

G.J Mysore [8] and A. Ozerov [10] presented the non-negative hidden Markov model (N-HMM), which jointly captures the spectral structure and temporal dynamics of a single source. This model uses several small dictionaries to capture the spectral structure of a sound source, in order to cater to the non-stationarity of audio. Additionally, a Markov chain is used to model the structure of changes between these dictionaries (temporal dynamics). Given a sound source, the dictionaries and the Markov chain are jointly learned.

Presentation of N-HMM

In the N-HMM, each of those dictionaries corresponds to one aspect of the sound, the transitions between the dictionaries represent the temporal dynamics. Those transitions are represented by a Markov chain and could be as well learned from the data. In this Markov chain, each state corresponds to one of the learned dictionaries.

Probabilistic Model

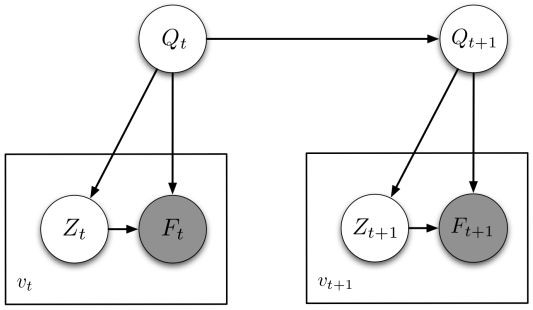

Figure 3: Graphical model for the N-HMM [8]

The figure 3 represents the transition between two states Q_t and Q_{t+1} For each state Q_t and Q_{t+1} exists a different dictionary. Each dictionary has several latent components, z which in our case would be spectral vectors. the spectral z of the stae q is define by the multinomial distribution: $P(f|z,q)$.

At a time frame t, only one of the state can be active and the data (spectrogram) in that time frame is modeled as a linear combination of the elements (z) of the corresponding dictionary. At time t, the weights of the different elements of dictionary are define by the multinomial distribution $P(z_t|q_t)$.

The temporal transitions from one state to an other is modeled by a transition matrix defined by $P(q_{t+1}|q_t)$ The different parameters can be estimated using EM algorithms as detailed in [8].

N-FHMM to Model Sound Mixtures

The N-HMM are useful to deal with single source audio files. The N-FHMM (non-negative factorial hidden markov models could be seen as an extension of N-HMM and permit to deal with sound mixtures in order, for example, to perform source separation. The figure 4 shows how N-FHMM combine N-HMM of different single sources. It introduces a new variable S_t, which indicates the weight of each source at each time frame.

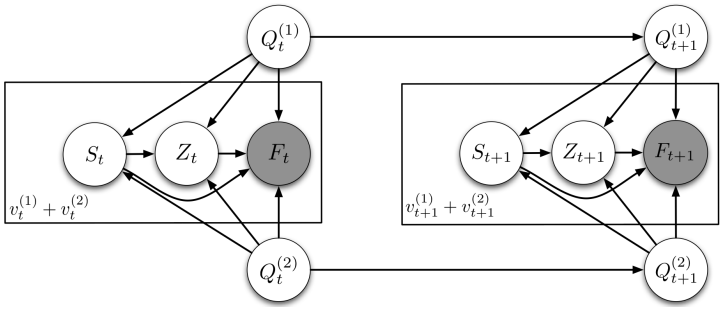

Figure 4: Graphical model for the N-FHMM [8]

In a given time frame t, each source is described thanks to one of its dictionary as seen for the N-HMM. Therefore, for a given mixture of n sources, each time frame is described using n dictionaries. In the particular case of a sound mixture of two sources, we have:

$$P(f_t|q_t^{(1)}, q_t^{(2)}) = \sum_{s_t} \sum_{z_t} P(f_t|z_t, s_t, q_t^{(1)}, q_t^{(2)}) \tag{1.15}$$

The mixture spectrum is modeled as a linear combination of the individual sources which are in turn modeled as a linear combination of spectral vectors from the given dictionaries. In the case of two sources, the mixture is a linear combination of the spectral vectors (z) of the given pair of dictionaries. The N-FHMM seems like an attractive method to capture temporality as, unlike the consultive NMF, it allows time stretching or modulation of the spectral components of the dictionary. However, it is an heavy method to implement and some problems such as to model the contribution of each source in the case of an unsupervised learning remain to be solved.

2 Application to source detection

2.1 Corpus and evaluation protocol

The aim of this experiment is to compare the performances achieved on sources detection by the different NMF algorithms we described in section 1 and to illustrate the influence of the parameters of the NMF on the achieve detection. It means that we focus on the ability of the algorithm to identify the time intervals during which a source is present without considering the quality of a possible reconstruction as we would have done in the case of source separation.

2.1.1 Description of the artificial scenes

Eight artificial 15 seconds scenes have been created in order to illustrate the application of the NMF on source detection. Four of those scenes are composed of drum sounds (kick, snare, tom and hat), chosen because of their low non-stationarity. The four others are constructed such as to be closer from what we would expect in a real-life auditory scene and are composed of bell ring, phone ring, footsteps, dog barking, diesel car engine, woman voice and the tune of the announcement in a French Train Station. All the sound sources come from mono files encoded at 44100 Hz. The scenes are as well mono encoded at 44100 Hz. Each scene is created by 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 when it is not.

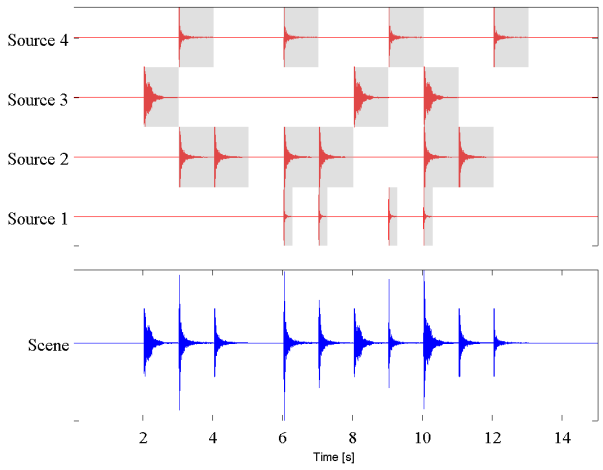

Figure 5: Construction of a test scene. The binary truth appears in grey background for each active source.

We will refer to the four scenes composed of drum sounds by D1, D2, D3, D4 and to the ones based on realistic sounds by R1, R2, R3 and R4. In order to evaluate the robustness of the algorithm, pink noise with an energy decreasing at 3dB per octave is added to each scene. We adjust the level of noise according to what we perceive as encountered in a quiet room and in an usual public space. According to that personal judgement, we apply a signal to noise ratio of 0.1 *Db* and 10 *Db*. We will refer to the scenes containing added noise by subscript indices. For example, for the scene D1, $D1_{0.1}$ and $D1_{10}$ refer respectively to the same scene with a SNR of 0.1 *Db* and 10 *Db*. D represent the entire group of Drum scenes.

2.1.2 Evaluation metric

The confusion matrix

The confusion matrix is a well know tool in order to evaluate the performance of a two groups classification task when provided with a score and a truth vector, both of same length. In our study, the classification task consists in labeling the sources as active or inactive for each sample of the input signal. In this process of evaluation, the confusion matrix can be established for each i^{th} source present in the scene and by comparing the corresponding activation coefficient, i^{th} line of $\mathbf{H}$ with the i^{th} line of the binary truth. Before doing so, we had to first rescale the binary truth constructed from the audio files to make it fit the length of the activations coefficients in $\mathbf{H}$. Let k be the difference between the length of the Hamming window and the length of the overlap used to compute the spectrogram, we have:

$$\forall i \in \mathbb{N}_{n_s} \qquad T_{fit}(i) = 1 \Leftrightarrow \frac{\sum_{(i-1)k+1}^{i\times k} T_i}{k} \geq 0.5 \tag{2.1}$$

For a given real value of threshold $\mathcal{T} \in \mathbb{R}^+$, the source is considered to be active at the sample i if $\mathbf{H}(i) \geq tr$.

Observation / *Truth*	Positive	Negative
Positive	a	b
Negative	c	d

Table 1: Confusion Matrix

a	is the number of true positives	b	is the number of false positives
c	is the number of false negatives	d	s the number of true negatives

$$recall = \frac{a}{a+b} \qquad precision = \frac{a}{a+c}$$

ROC Curve

We can establish the confusion matrix for different values of treshold $\mathcal{T}$. Then, for or each value of $\mathcal{T}$, the confusion matrix gives us a true-positive rate and a false positive rate. In our study, the true positive rate evaluates the ability of the algorithm to detect if a source is present or not. On the other hand, the false positive rate evaluates the tendency of the algorithm to produce false alarms.

As the value of $\mathcal{T}$ increases, the algorithm will classify more elements as positive, but may as well commit more false alarms. As both ratios can only take value in $[0, 1]$, the area under the ROC curve (*Receiver Operating Characteristic*) can be used as a good evaluation score that will take values in $[0, 1]$. This area under curve, that we note AUC, is used as the performance score in our source detection experiments. The ROC curves is also a good tool to visualize the performance of an algorithm for multiple values of $\mathcal{T}$. Indeed, a perfect classifier would be represented by a square and a worthless classifier would have its curve appearing under the diagonal of the graph.

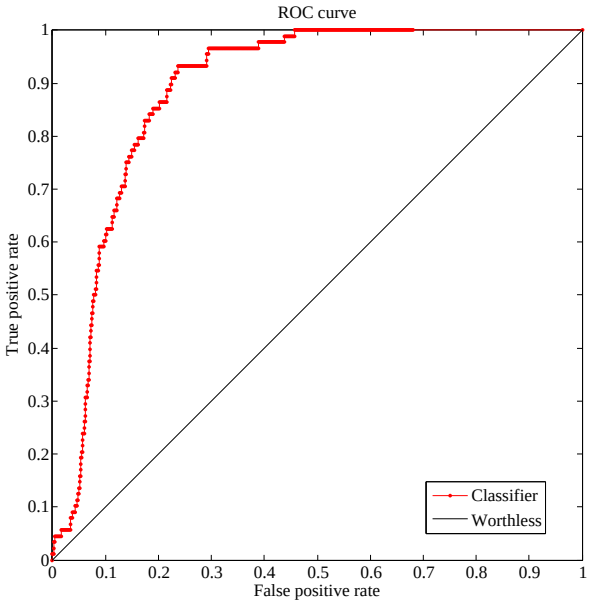

Figure 6: Example of the ROC curve obtained for the source detection applied on the scene D1 using supervised convolutive NMF with a one second length object and $sH_v = 0.8$.

Matching between H and $\mathbf{T}_{fit}$

Even though we always evaluate the source detection with the AUC, some distinctions have to be made in regard to the matching we establish between the time varying coefficient in $\mathbf{H}$ and the line of the binary truth $\mathbf{T}_{fit}$. In the case of supervised learning, as the dictionary is given as a prior, we know which of the sources is represented by each line of $\mathbf{H}$ and the corresponding line in $\mathbf{T}_{fit}$. Therefore, we simply calculate the AUC for each source and the global performance of the algorithm on one scene is the mean of the values obtained for each of its tracks.

However, in the case of an unsupervised learning we have no knowledge of the organization of $\mathbf{W}$ and $\mathbf{H}$ and do not know to which lines of $\mathbf{H}$ compare the lines of $\mathbf{T}_{fit}$. We established different metrics based on the AUC but that differ from each other by the used matching.

First, we can evaluate the AUC of each track for all the possible permutations of the lines of $\mathbf{H}$, and keep the best global score AUC_{bp} as the performance of our algorithm. However, the matching may not make correspond the time activations described in T_{fit} with their significant elements of dictionary if the source separation is not reliable enough. For example, it does not take into account that in NMF, a single element of dictionary can be used to reconstruct several of the sources. Therefore, this performance may provide artificially good results. Moreover, as the score has to be computed for each permutation, the computation time increases exponentially with the order r of the factorization and the AUC_{bp} cannot be applied to applications containing numerous sources.

Though the learning is unsupervised, the elements of dictionary $\mathbf{W}_s$used in the supervised learning algorithm can be used to guide the evaluation. First, the rows of $\mathbf{H}$ are sorted according to their level of energy and the same permutation is applied to the elements of the dictionary $\mathbf{W}$. $\mathbf{H}$ is then composed of r rows with $\mathbf{H}(r)$ the row of maximum energy. we evaluate the distances between $\mathbf{W}(r)$ and the different elements of $\mathbf{W}_s$ in order to identify the closest element $\mathbf{W}_s(i)$. The AUC is finally evaluated between $\mathbf{H}(r)$ and $\mathbf{T}_{fit}(i)$. This process is iterated along all lines of $\mathbf{H}$ by descending level of energy and the global performance is the mean of the respective values for each track. As this method is an oracle method based on the dictionaries $\mathbf{W}$, we note the resulting score AUC_{ow}. At each step of the iterative process, we could exclude the previously selected rows of $\mathbf{T}_{fit}$ as possible match for the currently evaluated row of $\mathbf{H}$. We will note the resulting score AUC_{owr}.

A similar method can be applied using $\mathbf{T}_{fit}$ as input of the oracle. In this case, after having sorted the rows of $\mathbf{H}$ according to their level of energy, each row is matched with the closest row of $\mathbf{T}_{fit}$. Again, we can choose to consider or not the previously selected rows of $\mathbf{T}_{fit}$ as possible match. The resulting score is noted AUC_{oh} when all matches remain possible and AUC_{ohr} when a line cannot be attributed twice.

2.2 Description of the experiment

On each of the artificial scenes, we applied the multiplicative update algorithm and the convolutive NMF algorithm. The scenes are 15 seconds mono *.wav* files encoded at 44100 Hz. The input of the algorithm is the spectrogram computed using the short-time Fourier transform with a Hamming window of length 1024 and an overlap of 50%.

As four sources have been added in each scene we set the order r of the factorization to $r = 4$ when no noise has been added and $r = 5$ when it has. In the case of the the convolutive NMF, the time length of the objects to be detected has been chosen in order to be superior to their actual length. It has been set to 1 second for the Drums scenes and to 2 seconds for the Realistic scenes. The vertical sparseness constraint sH_v 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. As $\mathbf{H}$, and $\mathbf{W}$ in the unsupervised case, are randomly initialized, ten run have been done with each set of parameters. The presented performances are the means of the score obtained for each of the ten runs. We present the values that illustrate best the influence of the parameters on the achieved source detection.

2.2.1 Supervised Source Detection

In the case of the supervised learning, a four elements dictionary has been input to the algorithm. Each element of dictionary has been learnt by applying the NMF algorithm with $r = 1$ and $sH_v = 0$ to the spectrogram of the audio file of each separated source. Two dictionaries have been built, corresponding to the multiplicative update algorithm and to the convolutive NMF algorithm. In the case of the convolutive algorithm, the length of each element of dictionary has been set to 1 second for the drum scenes and to 2 seconds for the realistic scene, which represent objects of 85 and 171 frames respectively.

Drum				
NMF	Multiplicative		Convolutive	
sH_v	**0**	**0.99**	**0**	**0.99**
AUC	0.87	0.92	0.82	0.95

Table 2: AUC of the supervised learning, multiplicative and convolutive, for $sH_v = 0$ and $sH_v = 0.99$, applied on the Drum scenes

Realistic				
NMF	Multiplicative		Convolutive	
sH_v	0	0.99	0	0.99
AUC	0.83	0.92	0.88	0.95

Table 3: AUC of the supervised learning, multiplicative and convolutive, for $sH_v = 0$ and $sH_v = 0.99$, applied on the Realistic scenes

We can see that even for a simple case, without added background noise, the convolutive NMF achieves better performances than the standard multiplicative update when applied to the group of the Realistic scenes. This improvement of performances does not appear for the Drum scenes. It comforts the hypothesis that convolutive NMF should improve the source detection of non-stationary sound.

2.2.2 Unsupervised Source Detection

In the case of the unsupervised detection, no dictionary has been input to the algorithm and therefore after having been randomly initialized, **W** is learnt and updated at each iteration along with **H**.

	sH_v	AUC_{bp}	AUC_{oh}	AUC_{ow}	AUC_{ohr}	AUC_{owr}
D	**0**	0.64	0.91	0.61	0.77	0.65
D	**0.99**	0.70	0.94	0.63	0.75	0.66
D_{10}	**0**	0.61	0.77	0.65	0.64	0.70
D_{10}	**0.99**	0.67	0.79	0.59	0.64	0.57
D_{01}	**0**	0.64	0.66	0.76	0.67	0.72
D_{01}	**0.99**	0.66	0.72	0.75	0.71	0.61

Table 4: The different types of AUC achieved on the Drum scenes by the convolutive NMF in function of sH_v and of the level of background noise

	sH_v	AUC_{bp}	AUC_{oh}	AUC_{ow}	AUC_{ohr}	AUC_{owr}
R	**0**	0.62	0.92	0.60	0.80	0.60
R	**0.99**	0.68	0.94	0.67	0.83	0.71
R_{10}	**0**	0.84	0.66	0.61	0.79	0.65
R_{10}	**0.99**	0.80	0.77	0.67	0.80	0.70
R_{01}	**0**	0.56	0.67	0.51	0.66	0.57
R_{01}	**0.99**	0.62	0.73	0.51	0.61	0.60

Table 5: The different types of AUC achieved on the Realistic scenes by the convolutive NMF in function of sH_v and of the level of background noise

As one may expect, the detection performance decreases when the signal to noise ratio increases. However, the sparseness constraint sH_v contributes to the robustness of the detection in presence of noise. We can also note that AUC_{oh} is higher than AUC_{bp}, which seems like each sources is represented by several elements of dictionary.

3 Application to Auditory Scenes Classification

3.1 Corpus and evaluation protocol

Our work is based on records collected by J. Tardieu [15] [14] in a study of the human perception of the similarity between soundscapes of train station. The corpus is composed of 66 audio files recorded in 6 different French train stations: Avignon TGV, Bordeaux St Jean, Lille Flandres, Nantes, Paris Gare de l'Est and Rennes. He considered that the typology of spaces in a train station is composed of six types: platform, hall, corridor / stair, waiting room, ticket office, shop. At least five recordings of about 3 minutes were made in each type of space and in each of the train stations for a total of nine hours during three days of recording sessions. Among those recording, a selection of 66 samples was made by four people. It permitted to remove the poor quality recordings and to select samples supposed to be representative of each space in terms of sound sources and human activity. In our categorization task, we will consider the 6 types of space as the six groups constituting of our ground truth.

J.Tardieu showed that the knowledge people have about train stations concerns the objects present in the spaces, the type of events happening and the type of space where all the sounds occur. Because of the ability of the NMF to isolate different sound sources and because of its additive nature that makes it an intuitive representation, we believe it is an interesting process to be used as a classification tool.

3.2 Evaluation

n-precion, Mean Average Precision and R-precision

The achieved classification has been evaluated using the n-precision at rank 5, the R-precision and the mean average precision.

$$\text{precision} = \frac{|\{\text{relevant samples}\} \cap \{\text{retrieved samples}\}|}{|\{\text{retrieved samples}\}|} \tag{3.1}$$

The precision at rank n, or *n-precision*, is the precision achieved for:
$$|\{\text{relevant samples}\} \cap \{\text{retrieved samples}\}| = n$$

The *r-precision* is the precision at rank r where $r = |\{\text{relevant samples}\}|$

The average precision AP is the average of the precision values at the points at which each relevant document is retrieved. The *Mean Average Precision* is the average of the average precision value for a set of queries.

$$MAP = \frac{\sum_{q=1}^{\mathcal{Q}} AP(q)}{\mathcal{Q}} \tag{3.2}$$

Recognition scores

The perceptual study gives the recognition scores achieved by the subjects during the forced recognition task. This recognition scores represent the average of the confusions matrix built for each subject during the experiment. In the computational method we consider, the computed distances between each of the samples produce a similarity matrix that we use to establish the recognition scores of each method. However, in order to do so, a K-Nearest Neighbors technique has to be applied and the recognition scores presented for the computational method depend of a chosen number of elements we consider in each group. We fixed this number at 11, which is the average number of elements per group in our considered corpus and means finding the 10 nearest neighbors for each sample.

For all methods, we define the recognition rate $\mathcal{R}$ as the mean of the diagonal of the recognition scores matrix. $\mathcal{R}$ is the mean of the percentages of correct attributed samples for each type of space of our ground truth. When compared to a random classifier, results are the mean of one hundred computation.

3.3 Previous classification methods

We do not aim here to present all existing classification methods used in computational auditory scene analysis but to describe the two methods with which compare our results. First, the perceptual study based on the human perception that we believe to be the top performance we aim to achieve. Then, the bag of frame approach, which is a computational method that has already been used on the classification of soundscapes, will be our baseline performance.

3.3.1 Perceptual study

This perceptual study was made by J.Tardieu on the same sound corpus we use. We describe here the six alternatives forced-choice recognition that we aim to reproduce with our computational method. Thirty-eight new people participated in the second experiment (17 women and 21 men, between 25 and 45 years old . The participants were provided with a description of the six types of spaces mentioned in 3.1. The experiment consisted of two steps, a six-alternative-forced-choice recognition task and a selection of prototypes. During the forced categorization the audio files were all randomly placed on one half of the screen and the participants had to sort the files into labeled categories the labeled categories corresponding to the types of the spaces where the samples were recorded. Participants could listen to each file as many times as they wanted.

For each type of space, it is possible to know the percentage of participants who attributed the samples to the right category.

	Associated Spaces					
Spaces	Platforms	Halls	Corridors	W. rooms	T. offices	Shops
Platforms	**67**	12	7	9	1	4
Halls	19	**52**	11	6	4	8
Corridors	16	14	**55**	2	5	8
W. rooms	6	7	12	**45**	9	21
T. offices	2	4	2	11	**53**	28
Shops	3	20	5	5	9	**57**

Table 6: Recognition scores for forced recognition Task: $\mathcal{R} = 54,8$

Except for the *Waiting room* (45%), the diagonal is greater than (50%). As there were six categories, a random choice would have resulted in $16,6\%$ in each cell of the table.

3.3.2 The "Bag of Frames" approach

Description

BOF models soundscapes as the long-term accumulative distribution of frame-based spectral features. The signal is cut into short overlapping frame typically 50 ms with 50% overlap. For each frame generic spectral features are computed (as MFCC, *Mel-frequency Cepstral Coefficients*). Then the spectral features are fed to a classifier (as GMM, *Gaussian Mixture Model*). The classifier models the global distribution of the features corresponding to each class. The global distribution for each class can then be used to determine decision boundaries between classes. A new, unobserved signal is classified by computing its features vectors, finding the most probable class for each of them. The most represented class is then the class of this new unobserved signal.

The BOF method has been used in order to classify soundscapes and musical pieces as well. J.J Aucouturier and F. Pachet studied the influence of the different parameters of this method in its application to the timbre similarity in [**?**]. The main algorithm can be summarize by this figure:

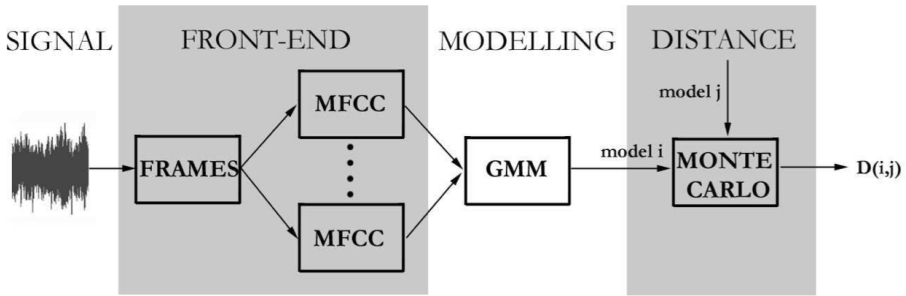

Figure 7: Bag Of Frame

Six parameters can influence the algorithm performance:

1 Signal Sample Rate (sr)
2 Number of MFCC (N_{mfcc})
3 Number of components (N_{gmm})
4 Distance Sample Rate (dsr)
5 Alternative Distance (ad)
6 Window Size (ws)

Study from J.J. Aucouturier [2]

This article is a comparison between the results given by the BOF approach to audio pattern recognition for Urban Soundscapes and polyphonic music. It concludes that the BOF approach is not a sufficient model to be applied to polyphonic music but that it does give convincing results on soundscapes classification in the case of urban soundscapes in spaces such as parks, avenues or market. However, in our particular corpus of train station soundscapes, the achieved performances are close from a random classifier.

	5-Precision	MAP	R-Precision
Random	0.18	0.25	0.19
Bag of Frames	0.18	0.24	0.20

Table 7: Achieved performance for the Bag of frame and for a random classifier

	Associated Spaces					
Spaces	Platforms	Halls	Corridors	W. rooms	T. offices	Shops
Platforms	**0**	30	30	20	20	0
Halls	12.5	**12.5**	62.5	0	12.5	0
Corridors	0	58.3	**25**	0	16,7	0
W. rooms	0	15.4	30.8	**15.4**	38.5	0
T. offices	0	0	30	10	**60**	0
Shops	20	0	80	0	0	**0**

Table 8: Recognition scores for the BOF method: $\mathcal{R} = 19,7$

3.4 Classification using NMF

3.4.1 Method

The 66 mono *.wav* files have been resample to 44100 Hz. Their spectrogram, computed as in 2.2, has been input to unsupervised NMF algorithms, multiplicative and convolutive in order to extract both $\mathbf{H}$ and $\mathbf{W}$. The rank R of the factorization has been set to 10, 25 and 50 elements of dictionary in the case of the multiplicative update algorithm. However, because of the high computation time required For each value of R, sH_v has been ranged to 0, 0.5, 0.8 and 0.99.In the case of the convolutive NMF, the time length of the objects as been set to a half second.
For each of the extraction, the distance between the different scenes have been computed in order to built the similarity matrix representing the corpus.

Measure of distance

We considered the distance between the scenes to be the distance between their respective dictionaries. In order to reduce the computation time and to have a representation closer from the human perception, the elements of dictionary have had their potential zero values replaced by their median before being reduced to 13 Mel Frequency Cepstral Coefficients (MFCC), which have then been normalized to take values between zero and one.

The distance between two elements from two different dictionaries W_a and W_b, after they have been normalized, is the sum of their element by element absolute difference. The distance between W_a and W_b is the sum of the R minimum distance between their elements of dictionary. Similarly to 2.1.2, while choosing the minimum pairwise distance between the elements of dictionary, we can consider than one element can be attributed only one time or not. Moreover, in order to determine wether or not the weight of each element of dictionary in the reconstruction would be a relevant feature to be used in the classification, the two distances have been calculated with and without a previous ponderation of each $\mathbf{W}(i)$ by $max(\mathbf{H}(i))$.

3.5 Classification experiments using NMF

3.5.1 Achieved performances

The multiplicative update is used in order to evaluate the influence of the different parameters. The highest scores have been reached for the distance computed by allowing multiple consideration of a single element of dictionary and without ponderation. In the following, we will consider the results achieved with that distance.

Multiplicative Update

The best classification performance achieved during our experiment has been achieved for the multiplicative update algorithm set with order of the NMF $R = 50$, a sparseness constraint $sH_v = 0.99$ and after 10 iterations.

	5-Precision	MAP	R-Precision
Random	0.18	0.25	0.19
Bag of Frames	0.18	0.24	0.20
Multiplicative NMF	0.45	0.31	0.22

Table 9: Achieved performances, Multiplicative NMF: $R = 50$, $sH_v = 0.99$

	Associated Spaces					
Spaces	Platforms	Halls	Corridors	W. rooms	T. offices	Shops
Platforms	**10**	60	10	0	20	0
Halls	0	**75**	18.7	0	6.2	0
Corridors	0	33.3	**66.7**	0	0	0
W. rooms	15.4	23	38.5	**15.4**	7.7	0
T. offices	0	0	30	30	**40**	0
Shops	0	0	100	0	0	**0**

Table 10: Recognition scores for the NMF, $\mathcal{R} = 34.8$ multiplicative update,$R = 50$, $sH_v = 0.99$

The achieved classification represents a significant improvement of the results achieved by the Bag of Frames approach.

Convolutive

Unfortunately, because of the long time of computation required by the convolutive NMF algorithm, we did not study it as deeply as the multiplicative update algorithm yet. The provided scores are the ones achieved for $R = 10$ and no constraint sH_v. The best performance has once again be achieved for 10 iterations.

	5-Precision	MAP	R-Precision
Random	0.18	0.25	0.19
Bag of Frames	0.18	0.24	0.20
Multiplicative NMF	0.45	0.31	0.22
Convolutive NMF	0.42	0.28	0.22

Table 11: Achieved performances, Convolutive NMF: $R = 10$, $sH_v = 0$

The given performance for the convolutive NMF are difficult to interpret because of the lack of results for other settings. Further experiments need to be run in order to conclude about the relevance of the convolutive NMF algorithm in this particular application.

3.5.2 Observations on the influence of the parameters

Vertical sparseness

The vertical sparseness constraint is used to reduced the number of elements of dictionary active at the same time. Our hypothesis was that it would contribute to make the elements of dictionary more representative of separated sources and improve the classification. It turns out that the sH_v is indeed relevant in the application to a classification task but only if the order R of the NMF has been set high enough.

R	**10**				**50**			
sHv	0	0.5	0.8	0.99	0	0.5	0.8	0.99
MAP	0.41	0.41	0.41	0.41	0.43	0.43	0.44	0.45
$\mathcal{R}$	**31.8**	**31.8**	**27.3**	**30.3**	**30.3**	**33.3**	**33.3**	**34.8**

Table 12: MAP and $\mathcal{R}$ achieved for R set to 10 and 50 elements of dictionary and a varying sH_v

Number of iterations

During the classification experiment, the results have been saved for the 1^{st}, 5^{th}, 10^{th} and 15^{th} iteration. It appears that from the 1^{st} to the 10^{th}, the achieved classification improves as the cost function $\mathcal{C}$ decreases. However, for all of our tests, we note that the classification score achieved at the 15^{th} iteration is poorer. It means that once a certain level of fitness of the factorization has been reached, the algorithm updates features irrelevant for the classification task. A too high number of iterations is therefore not only time consuming but also counterproductive. It is an important observation as it shows that the technic of the early stop [] would be necessary in applications of the algorithm.

Conclusion

We examined the application of NMF to sources detection before applying it to auditory scenes classification. We examined different evaluation tools to be applied to the classification or to the sources detection, as well as several ways of computing the distances between the extracted features of the NMF algorithm. We also observed the lack of correlation between the cost function of the NMF and the classification scores. The comparison between our achieved results and the bag of frames shows that NMF represents a significant improvement in the classification rate, which was the goal we were aiming at in the first place.

However, some work still need to be done. We did not deeply studied the influence of the different metrics we established and the results we achieved using the multiplicative update algorithm should be validated by more numerous runs. As we noticed that the best classification score has been achieved for the highest order of NMF we tried, it would be interesting to know if a higher order of factorization could still improve the results. Finally, we should determine if the convolutive update can improve the results we achieved so far by experimenting with same set of parameters we set for the multiplicative update.

References

[1] Jean-Julien Aucouturier and Boris Defreville. Judging the similarity of soundscapes does not require categorization: evidence from spliced stimuli. *Journal of the Acoustical Society of America*, 125(4):2155–2161, 2009.

[2] Jean-Julien Aucouturier, Boris Defreville, and François 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*, 122(2):881–891, 2007.

[3] Arnaud Dessein. Incremental multi-source recognition with non-negative matrix factorization. Master's thesis, Université Pierre et Marie Curie, Paris VI, 2009.

[4] Z. Ghahramani and M.I. Jordan. Factorial hidden markov models. *Machine learning*, 29(2):245–273, 1997.

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

[6] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

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

[8] Gautham J Mysore. *A Non-negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures*. PhD thesis, Stanford University, 2010.

[9] 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.

[10] A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden markov model for polyphonic audio representation and source separation. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 121–124. IEEE, 2009.

[11] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[12] M. Rajapakse, J. Tan, and J. Rajapakse. Color channel encoding with nmf for face recognition. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 3, pages 2007–2010. IEEE, 2004.

[13] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):1–12, 2007.

[14] J. Tardieu, P. Susini, and F. Poisson. Soundscape design in train stations: perceptual study of soundscapes. In *Proceedings of the CFA/DAGA (Joint French/German acoustical societies meeting). Strasbourg*, 2004.

[15] 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.

[16] T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*. Citeseer, 2006.

[17] B. Wang and M.D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *Proc. DMRN Summer Conf*, pages 23–24, 2005.

[18] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.