NMF with time-frequency activations to model non-stationary audio events

Romain Hennequin, Roland Badeau, Bertrand David

NMF with time-frequency activations to model non stationary audio events

Romain Hennequin, *Student Member, IEEE,* Roland Badeau, *Member, IEEE,* and Bertrand David, *Member, IEEE*

***Abstract*—Real world sounds often exhibit time-varying spectral shapes, as observed in the spectrogram of a harpsichord tone or that of a transition between two pronounced vowels. Whereas the standard Non-negative Matrix Factorization (NMF) assumes fixed spectral atoms, an extension is proposed where the temporal activations (coefficients of the decomposition on the spectral atom basis) become frequency dependent and follow a time-varying ARMA modeling. This extension can thus be interpreted with the help of a source/filter paradigm and is referred to as source/filter factorization. This factorization leads to an efficient single-atom decomposition for a single audio event with strong spectral variation (but with constant pitch). The new algorithm is tested on real audio data and shows promising results.**

***Index Terms*—music information retrieval, non-negative matrix factorization, unsupervised machine learning.**

I. Introduction

THE decomposition of audio signals in terms of elementary atoms has been a large field of research for years. As we usually encounter very dissimilar audio events (both in their spectral and temporal characteristics), the decomposition on a single basis (such as the Fourier basis) is generally not sufficient to accurately explain the content of a large class of signals. Sparse decomposition techniques [1] use a redundant dictionary of vectors (called atoms) and try to decompose a signal using few of them (much less than the dimension of the space): thus the signal can be accurately decomposed with few elements. When atoms are designed to fit the signal (for instance harmonic atoms for musical signals [2]), these elements become more meaningful, and then a supervised classification can be performed to cluster atoms corresponding to a real event in the signal [3]. These methods are quite powerful and give good results. However, since the dictionary is fixed, it must be designed to fit all possible signals, which is not achievable in practice. Recently methods of data *factorization* were proposed to simultaneously extract atoms from the signal and provide a decomposition on these atoms. These techniques that we call *factorization* make use of the natural redundancy of the signal, mimicking human cognition which utilizes this redundancy to understand visual and audio signals: Principal Component Analysis, Independent Component Analysis [4], [5], sparse coding [6] or NMF [7] have been introduced both to reduce the dimensionality and to explain the whole data set by a few meaningful elementary objects. Thanks to the non-negativity constraint, NMF is able to provide a meaningful representation of the data: applied to musical spectrograms it will hopefully decompose them into elementary notes or impulses. The technique is widely used in audio signal processing, with a number of applications such as automatic music transcription [8], [9], [10] and sound source separation [11], [12], [13].

However, the standard NMF is shown to be efficient when the elementary components (notes) of the analyzed sound are nearly stationary, i.e. when the envelope of the spectra of these components does not change over time. Nevertheless, in several situations, elementary components can be strongly non stationary. In this article, we will focus on timbral variability, i.e. variability of the spectral shape that we can find in plucked strings sounds or singing voice (sounds of different vowels present greatly dissimilar spectral shapes). However, we will not address pitch variability that is encountered in vibrato or prosody. In case of a noticeable spectral variability, the standard NMF will likely need several non-meaningful atoms to decompose a single event, which often leads to a necessary post-processing (to cluster the different parts of a single source [14]). To overcome this drawback, Smaragdis [15] proposes a shift-invariant extension of NMF in which time/frequency templates are factorized from the original data: each atom then corresponds to a time-frequency musical event able to include spectral variations over time. This method gives good results, but does not permit any variation between different occurrences of the same event (atom), its duration and spectral content evolution being fixed.

In this paper, an extension of NMF is proposed in which temporal activation becomes frequency dependent: it thus can be interpreted with the help of the classical source/filter paradigm as a source/filter factorization. Our method includes AutoRegressive Moving Average (ARMA) filters estimated from the data, associates a time-varying filter with each source and learns the sources (atoms) in a totally unsupervised way. this method presents some similarity with Durrieu's work [16], [17] in which a source/filter model is used in a NMF framework to extract the main melody of musical pieces. This model permits to efficiently take the strong spectral variations of the human voice into account.

However, our approach is quite different since, in opposition to Durrieu's work, sources are learnt, a time-varying filter is associated with each source and the class of filter we use is more standard.

In section II, we introduce the source/filter decomposition as an extension of NMF. In section III, we derive an iterative

R. Hennequin, R. Badeau and B. David are with the Institut Telecom, Telecom ParisTech, CNRS LTCI, 46, rue Barrault - 75634 Paris Cedex 13 - France (email: <forename>.<surname>@telecom-paristech.fr)

The research leading to this paper was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, DESAM, and by the Quaero Program, funded by OSEO, French State agency for innovation.

Manuscript received December 8, 2009

algorithm similar to those used for NMF to compute this decomposition. In section IV, we present experiments of source/filter decomposition of the spectrogram of three different sounds, and compare this decomposition to the standard NMF. Conclusions are drawn in section V.

II. Model

A. NMF and extension

Given an $F \times T$ non-negative matrix $\mathbf{V}$ and an integer R such that $FR + RT \ll FT$, NMF approximates $\mathbf{V}$ by the product of an $F \times R$ non-negative matrix $\mathbf{W}$ and an $R \times T$ non-negative matrix $\mathbf{H}$:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH} \quad \left(\text{i.e. } V_{ft} \approx \hat{V}_{ft} = \sum_{r=1}^{R} w_{fr} h_{rt}\right) \tag{1}$$

This approximation is generally quantified by a cost function $\mathcal{C}(\mathbf{V}, \mathbf{W}, \mathbf{H})$ to be minimized with respect to (wrt) $\mathbf{W}$ and $\mathbf{H}$. A common class of cost functions is designed element-wise:

$$\mathcal{C}(\mathbf{V}, \mathbf{W}, \mathbf{H}) = \sum_{f=1}^{F} \sum_{t=1}^{T} d(V_{ft} | \hat{V}_{ft})$$

where d is a scalar divergence (i.e. a function such that $\forall (a,b) \in \mathbb{R}^2$ $d(a|b) \geq 0$ and $d(a|b) = 0$ if and only if $a = b$). Several classes of such divergences have been proposed, for instance Bregman divergences [18] and β-divergences [19], [20]. In this paper, we will focus on the β-divergence which includes usual measures (Euclidean distance, Kullback-Lieber divergence and Itakura-Saito divergence). The β-divergence is defined for $\beta \in \mathbb{R} \backslash \{0, 1\}$ by:

$$d_\beta(x|y) = \frac{1}{\beta(\beta-1)} (x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}) \tag{2}$$

For $\beta = 0$ and $\beta = 1$, the β-divergence is defined by continuity:

$$d_0(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1$$
$$d_1(x|y) = x(\log x - \log y) + (y - x)$$

For these values, the β-divergence respectively corresponds to Itakura-Saito divergence and Kullback-Leibler divergence, and for $\beta = 2$, it corresponds to the Euclidean distance.

One could notice that the singularities in the definition of the β-divergence for $\beta = 0$ and $\beta = 1$ no longer appear in the partial derivative wrt y: this partial derivative is useful for designing descent methods in order to minimize the cost function $\mathcal{C}$. The first order partial derivative of the β-divergence wrt y is for all $\beta \in \mathbb{R}$:

$$\frac{\partial d_\beta(x|y)}{\partial y} = y^{\beta-2}(y - x)$$

When applied to power (squared magnitude) spectrograms, NMF factorizes data into a matrix (or basis) of frequency templates which are the R columns of $\mathbf{W}$ and a matrix $\mathbf{H}$ whose R rows are the temporal vectors of activations corresponding to each template. For a musical signal made of several notes played by the same instrument, it is hoped that the decomposition leads to spectral templates corresponding to single notes or percussive sounds. $\mathbf{H}$ will then display a representation similar to a "piano-roll" (cf. [8]).

This factorization however does not yield an effective representation of a sound presenting a noticeable spectral evolution. For instance a single note of a plucked string instrument most of the time shows high frequency components which decrease faster than low frequency components. This characteristic is not well modeled with a single frequency template. Several templates are needed which results in a less meaningful decomposition: roughly one for low frequency partials and one for high frequency partials. The meaning of each template is lost (a template no longer corresponds to a musical event such as a note).

To address this issue, we propose an extension of NMF where temporal activations become time/frequency activations. The factorization (1) becomes:

$$V_{ft} \approx \hat{V}_{ft} = \sum_{r=1}^{R} w_{fr} h_{rt}(f) \tag{3}$$

where the activation coefficients are now frequency dependent. To avoid an increase of the problem dimensionality the $h_{rt}(f)$ coefficients are further parameterized by means of ARMA models (section II-B).

Equation (3) can be interpreted with the help of the source/filter paradigm: the spectrum of each frame of the signal results from the combination of filtered templates (sources). $h_{rt}(f)$ corresponds to the time-varying filter associated to the source r. The decomposition thus benefits from the versatility of the source/filter model proved well suited for numerous sound objects.

B. AutoRegressive Moving Average (ARMA) Modeling

$h_{rt}(f)$ is parameterized following the general ARMA model:

$$h_{rt}^{ARMA}(f) = \sigma_{rt}^2 \frac{\left| \sum_{q=0}^{Q} b_{rt}^q e^{-i2\pi\nu_f q} \right|^2}{\left| \sum_{p=0}^{P} a_{rt}^p e^{-i2\pi\nu_f p} \right|^2}$$

where $\nu_f = \frac{f-1}{2(F-1)}$ is the normalized frequency associated to frequency index $f \in \{1, ..., F\}$ (as audio signal are real valued, we only consider frequencies between 0 and the Nyquist frequency). b_{rt}^q are the coefficients of the MA part of the filter and a_{rt}^p those of the AR part. σ_{rt}^2 is the global gain of the filter: in order to avoid identifiability problems, the first coefficient of all filters is imposed to be equal to 1. For $P = Q = 0$, $h_{rt}^{ARMA}(f)$ no longer depends on f and the decomposition corresponds to a standard NMF with temporal activations σ_{rt}^2.

Defining $\mathbf{a}_{rt} = (a_{rt}^0, \ldots, a_{rt}^P)^T$ and $\mathbf{b}_{rt} = (b_{rt}^0, \ldots, b_{rt}^Q)^T$, time/frequency activations can be rewritten as:

$$h_{rt}^{ARMA}(f) = \sigma_{rt}^2 \frac{\mathbf{b}_{rt}^T \mathbf{T}(\nu_f) \mathbf{b}_{rt}}{\mathbf{a}_{rt}^T \mathbf{U}(\nu_f) \mathbf{a}_{rt}}$$

where $\mathbf{T}(\nu)$ is the $(Q+1)\times(Q+1)$ Toeplitz matrix with $[\mathbf{T}(\nu)]_{pq}=\cos(2\pi\nu(p-q))$ and $\mathbf{U}(\nu)$ is similar to $\mathbf{T}(\nu)$ but of dimension $(P+1)\times(P+1)$. MA only and AR only models are included by respectively taking $P=0$ and $Q=0$. It is worth noting that $h_{rt}^{ARMA}(f)$ is always non-negative while there exists no non-negativity constraint on $\mathbf{b}_{rt}^{q}$ or on $\mathbf{a}_{rt}^{p}$.

The parameterized power spectrogram given in equation (3) then becomes:

$$\hat{V}_{ft}=\sum_{r=1}^{R} w_{fr}\sigma_{rt}^{2}\frac{\mathbf{b}_{rt}^{T}\mathbf{T}(\nu_f)\mathbf{b}_{rt}}{\mathbf{a}_{rt}^{T}\mathbf{U}(\nu_f)\mathbf{a}_{rt}} \tag{4}$$

III. Algorithm

We choose a general β-divergence cost function:

$$\mathcal{C}(\mathbf{W},\mathbf{A},\mathbf{B},\mathbf{\Sigma})=\sum_{f=1}^{F}\sum_{t=1}^{T} d_\beta(V_{ft},\hat{V}_{ft})$$

with $[\mathbf{W}]_{fr}=w_{fr}$, $[\mathbf{\Sigma}]_{rt}=\sigma_{rt}^2$, $[\mathbf{A}]_{rtp}=a_{rt}^{p}$ and $[\mathbf{B}]_{rtq}=b_{rt}^{q}$ and the expression of d_β is given in equation (2).

The partial derivative of the cost function wrt any variable θ (θ being any coefficient of $\mathbf{W}$, $\mathbf{\Sigma}$, $\mathbf{A}$ or $\mathbf{B}$) is:

$$\frac{\partial\mathcal{C}(\mathbf{W},\mathbf{A},\mathbf{B},\mathbf{\Sigma})}{\partial\theta}=\sum_{f=1}^{F}\sum_{t=1}^{T}\hat{V}_{ft}^{\beta-2}(\hat{V}_{ft}-V_{ft})\frac{\partial\hat{V}_{ft}}{\partial\theta} \tag{5}$$

The expression of the gradient of $\mathcal{C}$ wrt a vector θ of several coefficients of $\mathbf{A}$ or $\mathbf{B}$ is the same, replacing the partial derivative by a gradient ∇_θ in (5).

This leads to update rules for a multiplicative gradient descent algorithm similar to those used in [7], [21], [15]. In such an iterative algorithm, the update rule associated to one of the parameters is obtained from the partial derivative of the cost function wrt this parameter, written as a difference of two positive terms: $\frac{\partial\mathcal{C}}{\partial\theta}=G_\theta-F_\theta$

The update rule for θ is then:

$$\theta\leftarrow\theta\times\frac{F_\theta}{G_\theta} \tag{6}$$

This rule ensures that θ remains non-negative, becomes constant if the partial derivative is zero and evolves in the opposite direction of the partial derivative (thus in the descent direction).

A. Update of frequency templates:

We derive multiplicative update rules for w_{fr} from the expression of the partial derivative of the cost function with respect to w_{fr}.

The partial derivative of the parameterized spectrogram defined in equation (4) with respect to $w_{f_0r_0}$ is:

$$\frac{\partial\hat{V}_{ft}}{\partial w_{f_0r_0}}=h_{r_0t}^{ARMA}(f_0)\delta_{f_0f}$$

where δ_{f_0f} is a Kronecker delta.

Then, by replacing this expression in equation (5) with $\theta=w_{fr_0}$, we obtain the partial derivative of the cost function with respect to $w_{f_0r_0}$:

$$\frac{\partial\mathcal{C}(\mathbf{W},\mathbf{A},\mathbf{B},\mathbf{\Sigma})}{\partial w_{f_0r_0}}=\sum_{t=1}^{N}h_{r_0t}^{ARMA}(f_0)\hat{V}_{f_0t}^{\beta-2}(\hat{V}_{f_0t}-V_{f_0t})$$

This derivative is written as a difference of two positive terms:

$$G_{w_{f_0r_0}}=\sum_{t=1}^{N}h_{r_0t}^{ARMA}(f_0)\hat{V}_{f_0t}^{\beta-1}$$

and

$$F_{w_{f_0r_0}}=\sum_{t=1}^{N}h_{r_0t}^{ARMA}(f_0)\hat{V}_{f_0t}^{\beta-2}V_{f_0t}$$

Then the update rule of $w_{f_0r_0}$ is:

$$w_{f_0r_0}\leftarrow w_{f_0r_0}\frac{F_{w_{f_0r_0}}}{G_{w_{f_0r_0}}} \tag{7}$$

B. Update of temporal activation gain:

In the same way as for w_{fr}, we derive multiplicative update rules for σ_{rt}^2 from the expression of the partial derivative of the cost function with respect to σ_{rt}^2.

The partial derivative of the parameterized spectrogram defined in equation (4) with respect to $\sigma_{r_0t_0}^2$ is:

$$\frac{\partial\hat{V}_{ft}}{\partial\sigma_{r_0t_0}^2}=w_{fr_0}\frac{\mathbf{b}_{r_0t_0}^{T}\mathbf{T}(\nu_f)\mathbf{b}_{r_0t_0}}{\mathbf{a}_{r_0t_0}^{T}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0}}\delta_{t_0t}$$

where δ_{t_0t} is a Kronecker delta.

Then, by substituting this expression into equation (5) with $\theta=\sigma_{r_0t_0}^2$, we obtain the partial derivative of the cost function with respect to $\sigma_{r_0t_0}^2$:

$$\frac{\partial\mathcal{C}(\mathbf{W},\mathbf{A},\mathbf{B},\mathbf{\Sigma})}{\partial\sigma_{r_0t_0}^2}=\sum_{f=1}^{M}\frac{w_{fr_0}}{\sigma_{r_0t_0}^2}h_{r_0t_0}^{ARMA}(f)\hat{V}_{ft_0}^{\beta-2}(\hat{V}_{ft_0}-V_{ft_0})$$

This derivative is written as a difference of two positive terms:

$$G_{\sigma_{r_0t_0}^2}=\sum_{f=1}^{M}\frac{w_{fr_0}}{\sigma_{r_0t_0}^2}h_{r_0t_0}^{ARMA}(f)\hat{V}_{ft_0}^{\beta-1}$$

and

$$F_{\sigma_{r_0t_0}^2}=\sum_{f=1}^{M}\frac{w_{fr_0}}{\sigma_{r_0t_0}^2}h_{r_0t_0}^{ARMA}(f)\hat{V}_{ft_0}^{\beta-2}V_{ft_0}$$

Then the update rule of $\sigma_{r_0t_0}^2$ is:

$$\sigma_{r_0t_0}^2\leftarrow\sigma_{r_0t_0}^2\frac{F_{\sigma_{r_0t_0}^2}}{G_{\sigma_{r_0t_0}^2}} \tag{8}$$

We can notice that when $Q=0$ and $P=0$ (i.e. there is no filter), the update rules of w_{fr} and σ_{rt}^2 are the same as the ones given in [21] which are the standard NMF rules for a β-divergence cost function where w_{fr} corresponds to frequency templates and σ_{rt}^2 to temporal activations.

C. Update of filters

The update rules of the coefficients of the filters are derived in a similar way, but the updates are not element-wise, but rather "vector-wise": we derive an update rule for each $\mathbf{b}_{rt}$ and for each $\mathbf{a}_{rt}$.

Update of $\mathbf{b}_{rt}$: The gradient of the parameterized spectrogram $\hat{V}_{ft}$ wrt $\mathbf{b}_{r_0t_0}$ is:

$$\nabla_{\mathbf{b}_{r_0t_0}} \hat{V}_{ft} = \delta_{t_0t} \frac{2w_{fr_0}\sigma^2_{r_0t_0}}{\mathbf{a}^T_{r_0t_0}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0}} \mathbf{T}(\nu_f)\mathbf{b}_{r_0t_0}$$

Then, by substituting this expression into equation (5) with $\theta = \mathbf{b}_{r_0t_0}$, we obtain the gradient of the cost function wrt $\mathbf{b}_{r_0t_0}$:

$$\begin{aligned}
\nabla_{\mathbf{b}_{r_0t_0}} \mathcal{C} &= 2\sum_{f=1}^{F} \frac{w_{fr_0}\sigma^2_{r_0t_0}\hat{V}^{\beta-2}_{ft_0}(\hat{V}_{ft_0} - V_{ft_0})}{\mathbf{a}^T_{r_0t_0}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0}} \mathbf{T}(\nu_f)\mathbf{b}_{r_0t_0} \\
&= 2\sigma^2_{r_0t_0}\left(\mathbf{R}_{r_0t_0} - \mathbf{R}'_{r_0t_0}\right)\mathbf{b}_{r_0t_0}
\end{aligned}$$

where:

$$\mathbf{R}_{r_0t_0} = \sum_{f=1}^{F} \frac{w_{fr_0}\hat{V}^{\beta-1}_{ft_0}}{\mathbf{a}^T_{r_0t_0}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0}} \mathbf{T}(\nu_f)$$

$$\mathbf{R}'_{r_0t_0} = \sum_{f=1}^{F} \frac{w_{fr_0}\hat{V}^{\beta-2}_{ft_0}V_{ft_0}}{\mathbf{a}^T_{r_0t_0}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0}} \mathbf{T}(\nu_f)$$

Both matrices $\mathbf{R}_{r_0t_0}$ and $\mathbf{R}'_{r_0t_0}$ are positive definite under mild assumptions: these matrices are clearly positive semi-definite and it can easily be shown that if there are at least $Q+1$ different indexes f such that $w_{fr_0}\hat{V}^{\beta}_{ft_0} \neq 0$ then $R_{r_0t_0}$ is non-singular (for $R'_{r_0t_0}$, the assumption is very similar). This assumption is always true in practice as long as the frame with index t_0 is not a null vector (*i.e.* with all samples equal to 0): in this particular case, the decomposition is trivial, and the global gains $\sigma_{r_0t_0}$ should be equal to 0.

Then, we follow the approach given in [22] and derive the following update rule for the MA part of the filter:

$$\mathbf{b}_{r_0t_0} \leftarrow \mathbf{R}^{-1}_{r_0t_0}\mathbf{R}'_{r_0t_0}\mathbf{b}_{r_0t_0} \tag{9}$$

As $\mathbf{R}_{r_0t_0}$ and $\mathbf{R}'_{r_0t_0}$ are both non singular, $\mathbf{R}^{-1}_{r_0t_0}$ is well defined and $\mathbf{b}_{r_0t_0}$ is ensured to never be zero.

Update of $\mathbf{a}_{rt}$:

The update rules of $\mathbf{a}_{rt}$ are derived in the same way as for $\mathbf{b}_{rt}$. The partial gradient of the parameterized spectrogram $\hat{V}_{ft}$ with respect to $\mathbf{a}_{r_0t_0}$ is:

$$\nabla_{\mathbf{a}_{r_0t_0}} \hat{V}_{ft} = -2\delta_{tt_0} w_{fr_0} \frac{h^{ARMA}_{r_0t_0}(f)}{\mathbf{a}^T_{r_0t_0}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0}} \mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0}$$

Then, by substituting this expression into equation (5) with $\theta = \mathbf{a}_{r_0t_0}$, we obtain the partial gradient of the cost function with respect to $\mathbf{a}_{r_0t_0}$:

$$\nabla_{\mathbf{a}_{r_0t_0}} \mathcal{C}(\mathbf{W}, \mathbf{A}, \mathbf{B}, \mathbf{\Sigma}) = 2\sigma^2_{r_0t_0}\left(\mathbf{S}'_{r_0t_0} - \mathbf{S}_{r_0t_0}\right)\mathbf{a}_{r_0t_0}$$

where:

$$\mathbf{S}_{r_0t_0} = \sum_{f=1}^{M} w_{fr_0}\hat{V}^{\beta-1}_{ft_0} \frac{\mathbf{b}^T_{r_0t_0}\mathbf{T}(\nu_f)\mathbf{b}_{r_0t_0}}{(\mathbf{a}^T_{r_0t_0}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0})^2} \mathbf{U}(\nu_f)$$

and

$$\mathbf{S}'_{r_0t_0} = \sum_{f=1}^{M} w_{fr_0}\hat{V}^{\beta-2}_{ft_0}V_{ft_0} \frac{\mathbf{b}^T_{r_0t_0}\mathbf{T}(\nu_f)\mathbf{b}_{r_0t_0}}{(\mathbf{a}^T_{r_0t_0}\mathbf{U}(\nu_f)\mathbf{a}_{r_0t_0})^2} \mathbf{U}(\nu_f)$$

Both matrices $\mathbf{S}_{r_0t_0}$ and $\mathbf{S}'_{r_0t_0}$ are positive definite under mild assumptions.

Thus we derive the following update rule for the AR part of the filter:

$$\mathbf{a}_{r_0t_0} \leftarrow \mathbf{S}'^{-1}_{r_0t_0}\mathbf{S}_{r_0t_0}\mathbf{a}_{r_0t_0} \tag{10}$$

D. Description of the algorithm

The update rules (7), (8), (9) and (10) are applied successively to all the coefficients of $\mathbf{W}$, all the coefficients of $\mathbf{\Sigma}$, all the coefficients of $\mathbf{B}$ and all the coefficients of $\mathbf{A}$. Between the updates of each of these matrices (and tensors), the parameterized spectrogram $\hat{\mathbf{V}}$ is recomputed: as for the standard NMF algorithm, this recomputation between each update is necessary to ensure the convergence.

Identifiability: As for the standard NMF, the decomposition (4) which minimizes the cost function is not unique. To cope with identifiability issues, we impose constraints on $\mathbf{W}$, $\mathbf{\Sigma}$, $\mathbf{B}$ and $\mathbf{A}$:

- for all r and t, we impose that $\mathbf{b}_{rt}$ and $\mathbf{a}_{rt}$ (considered as polynomials) have all their roots inside the unit circle.
- for all r, we impose $\|\mathbf{w}_r\| = 1$ for some norm $\|.\|$.
- for all r and t, we impose $b^0_{rt} = 1$ and $a^0_{rt} = 1$.

Thus, at the end of each iteration of our algorithm, we transform $\mathbf{b}_{r,t}$ and $\mathbf{a}_{r,t}$ by replacing roots outside the unit circle by the conjugate of their inverse and accordingly adapting the gain, normalize each column of $\mathbf{W}$, divide $\mathbf{b}_{r,t}$ and $\mathbf{a}_{r,t}$ by their first coefficient and update $\mathbf{\Sigma}$ in order not to change $\hat{V}_{ft}$ by these modifications. All these transformations have no influence on the values of the parameterized spectrogram.

Another choice of filters normalization has been tested: rather than imposing $b^0_{rt} = 1$ and $a^0_{rt} = 1$, we can impose for all (r,t), $\sum_f \mathbf{b}^T_{rt}\mathbf{T}(\nu_f)\mathbf{b}_{rt} = 1$ and $\sum_f \mathbf{a}^T_{rt}\mathbf{U}(\nu_f)\mathbf{a}_{rt} = 1$. It corresponds to a power normalization and then it is more meaningful.

Our algorithm is detailed in Algorithm 1. In the remainder of the article, we will refer to this algorithm by the expression "source/filter factorization".

E. Dimensionality

Since our algorithm is a dimensionality reduction technique like NMF, one should take care of the dimension of the decomposition provided. The dimension of the original data is FT. In the standard NMF with R atoms, the dimension of the parameterized spectrogram is $\dim \mathbf{W} + \dim \mathbf{H} = R(F+T)$ (where $\dim \mathbf{\Phi}$ stands for the number of coefficients of matrix or tensor $\mathbf{\Phi}$). With our algorithm, the dimension of the parameters is: $\dim \mathbf{W} + \dim \mathbf{\Sigma} + \dim \mathbf{A} + \dim \mathbf{B} = RF + RT(P+Q+1)$. Thus, one should have $RF + RT(P+Q+1) \ll FT$, i.e.

Algorithm 1 Source/filter spectrogram factorization

Input: $\mathbf{V}$, R, Q, P, n_{iter}, β
Output: $\mathbf{W}$, $\mathbf{\Sigma}$, $\mathbf{B}$, $\mathbf{A}$
  Initialize $\mathbf{W}$, $\mathbf{\Sigma}$ with non-negative values
  Initialize $\mathbf{B}$, $\mathbf{A}$ with flat filters
  for j = 1 to n_{iter} **do**
    compute $\hat{\mathbf{V}}$
    for all f and r **do**
      compute $F_{w_{fr}}$ and $G_{w_{fr}}$
      $w_{fr} \leftarrow w_{fr} \frac{F_{w_{fr}}}{G_{w_{fr}}}$
    end for
    compute $\hat{\mathbf{V}}$
    for all r and t **do**
      compute $F_{\sigma^2_{rt}}$ and $G_{\sigma^2_{rt}}$
      $\sigma^2_{rt} \leftarrow \sigma^2_{rt} \frac{F_{\sigma^2_{rt}}}{G_{\sigma^2_{rt}}}$
    end for
    compute $\hat{\mathbf{V}}$
    for all r and t **do**
      compute $\mathbf{R}_{rt}$, $\mathbf{R}'_{rt}$, $\mathbf{S}_{rt}$ and $\mathbf{S}'_{rt}$
      $\mathbf{b}_{rt} \leftarrow \mathbf{R}_{rt}^{-1}\mathbf{R}'_{rt}\mathbf{b}_{rt}$
      $\mathbf{a}_{rt} \leftarrow \mathbf{S}_{rt}'^{-1}\mathbf{S}_{rt}\mathbf{a}_{rt}$
    end for
    bring back roots of all filters inside the unit circle
    divide the coefficients of all filters by the first one
    normalize $\mathbf{W}$
    update appropriately $\mathbf{\Sigma}$
  end for

$R \ll T$ and $R(P+Q+1) \ll F$, so P and Q must be small. As in practice $F \leq T$, the condition to be respected is $R(P+Q+1) \ll F$.

One should notice that our decomposition allows a significant reduction of the number of atoms R needed to accurately fit the data when the parts of the sounds present strong spectral variations. Then the total dimension of the parameters obtained with our decomposition remains comparable to the one obtained with standard NMF as will be shown in section IV.

Besides, one should notice that a large number of the coefficients of the filters are useless and therefore do not need to be retained: when the global gain of one of the filter σ_{rt} becomes close to zero, these coefficients ($\mathbf{b}_{rt}$ and $\mathbf{a}_{rt}$) become meaningless and then are useless in the decomposition and can be removed without affecting the values of the parameterized spectrogram.

Finally, in the decomposition (4), all atoms are associated to filters of the same order, but it is also possible to implement a larger model where filters do not have the same characteristics for all atoms. This larger model is not presented in this paper for readability reasons.

F. Computational complexity

The computational complexity of one iteration of source/filter factorization depends on P, Q, R, F and T. The computational complexity of each step of the algorithm is given here:

- Computation of $\hat{\mathbf{V}}$: $\mathcal{O}\left((P+Q)\,RFT\right)$ operations.
- Update of $\mathbf{W}$ and $\mathbf{\Sigma}$: $\mathcal{O}\left(RFT\right)$ operations each.
- Update of $\mathbf{B}$: $\mathcal{O}\left(RT(FP+P^3)\right)$ operations.
- Update of $\mathbf{A}$: $\mathcal{O}\left(RT(FQ+Q^3)\right)$ operations.
- Normalization/stabilization: $\mathcal{O}\left(RT(F+P^3+Q^3)\right)$ operations.

The total complexity of a single iteration of the algorithm is then $\mathcal{O}\left((P+Q)\,RFT\right)$ operations. With our current implementation in Matlab, 100 iterations of our algorithm applied to a 1025×550 spectrogram (corresponding to a 6.5s signal sampled at $f_s = 22050$Hz with 2048-sample-long windows and 75% overlap) with $R = 10$ atoms, $P = 2$ and $Q = 2$ last about 300s (on an Intel®Core™2 Duo E8400 @3.00GHz, with Matlab's multithreaded math libraries). In comparison, 100 iterations of standard NMF with the same spectrogram, the same parameters ($R = 10$, but $P = 0$ and $Q = 0$) on the same computer last about 9s: our algorithm appears to be slower than standard NMF. However, this comparison puts at a disadvantage our algorithm, which is designed to work with fewer atoms than standard NMF: in this case our algorithm deals with many more parameters than NMF. If we compare execution times with the same number of parameters, the difference is smaller: for the same spectrogram, 100 iterations of our algorithm with $R = 2$ atoms, $P = 2$ and $Q = 2$ (i.e. with the dimensionality of a standard NMF with $R = 10$) last about 60s.

About a third of the computation time is due to the computation of the roots of all filters (considered as polynomials) during the stabilization process. Some improvements could be made by considering another regularization method. The inversion of the matrices R_{rt} and S'_{rt} (a bit more than 10% of the total computation time) and the computation the frequency response of all filters (slighly less than 10% of the total computation time) are also very costly.

G. Practical implementation and choice of β

We empirically observed the monotonic decrease of the cost function and the convergence of the algorithm for $0 \leq \beta \leq 2$ over a large set of tests: this decrease and this convergence are illustrated in particular cases in section IV-D.

However, the algorithm is unstable for $1 \leq \beta \leq 2$: some numerical instabilities appear while poles of the filters come close to the unit circle. These instabilities are very common when β becomes close to 2 (Euclidean distance); however, the strong dynamics of audio data is better fitted when β becomes close to 0 as stated in [21]. In order to avoid these instabilities, we limit the modulus of poles: the monotonic decrease of the cost function is no longer observed but it permits to avoid non desirable behavior of the decomposition (very resonant order 2 filters only permit to fit one partial).

For the examples in the next section, we chose $\beta = 0.5$ since the numerical instabilities were almost negligible and the results were more accurate than with $\beta = 0$ (Itakura-Saito divergence).

IV. EXAMPLES

In this section several experiments are presented to show that our algorithm is well adapted to decompose sounds having

strong spectral variations. All the spectrograms used in these experiments are power spectrograms obtained from recorded signals by means of a short time Fourier transform (STFT).

Algorithms (standard NMF and source/filter factorization) were initialized with random values (except for the filters which were initially flat) and were run until apparent convergence. The algorithms have been rerun with 100 different initializations in order to maximize the chances to come close to a "good" minimum. Despite these different starting points, the reached solutions were similar in terms of qualitative aspect of the reconstructed spectrograms.

A. Didgeridoo

1) Description of the excerpt: In this section our algorithm is applied to a short didgeridoo excerpt. The didgeridoo is an ethnic wind instrument from northern Australia. It makes a continuous modulated sound produced by the vibrations of the lips. The modulations result from the mouth and throat configuration with the help of which the player is able to control several resonances. Figure 1(a) represents the spectrogram of the excerpt: the sound produced is almost harmonic (with some noise) and a strong moving resonance appears in the spectrogram. We can thus consider that this signal is composed of a single event encompassing spectral variations, and try to decompose it with a single atom ($R = 1$). The sampling rate of the excerpt is $f_s = 11025Hz$. We chose a 1024-sample-long Hann window with 75% overlap for the STFT.

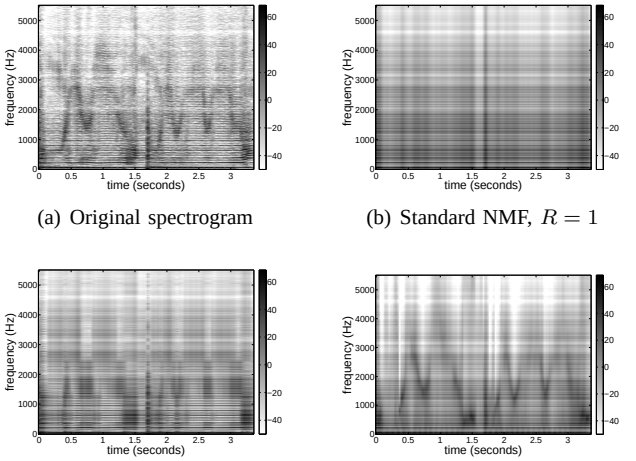
2) Experiment and results: The spectrogram of the excerpt is decomposed with a standard NMF algorithm for $R = 1$ atom and $R = 5$ atoms, and with source/filter factorization for $R = 1$ atom, with an order 3 AR modeling ($Q = 0$ and $P = 3$). Reconstructed spectrograms are respectively represented in figures 1(b), 1(c) and 1(d).

Although the didgeridoo is played alone in the analyzed spectrogram, the standard NMF needs many atoms to accurately decompose the power spectrogram. With 1 atom, NMF does not accurately represent the moving resonance (figure 1(b)). With 5 atoms, some spectral variations appear (figure 1(c)), but the resonance trajectory remains a bit unclear. Besides, the signal is not decomposed in a meaningful way (each atom is a part of the sound which has no perceptual meaning) and the dimensionality of the parameters is large ($FR + RT = 3290$).

In opposition to the standard NMF, source/filter factorization permits to accurately represent the spectral variability of the sound (figure 1(d)) with a single atom, keeping the dimensionality low ($FR + TR(Q + 1) = 1093$): the moving resonance of the original sound is well tracked, and the total error $\mathcal{C}$ is smaller than that of the standard NMF with $R = 5$. In this case, the decomposition is more efficient and relevant than the standard NMF.

B. Harpsichord

1) Description of the excerpt: In this section our algorithm is applied to a short harpsichord excerpt, composed of two different notes ($C2$ and $E\flat2$): first, the $C2$ is played alone, then the $E\flat2$, and at last, both notes are played simultaneously.

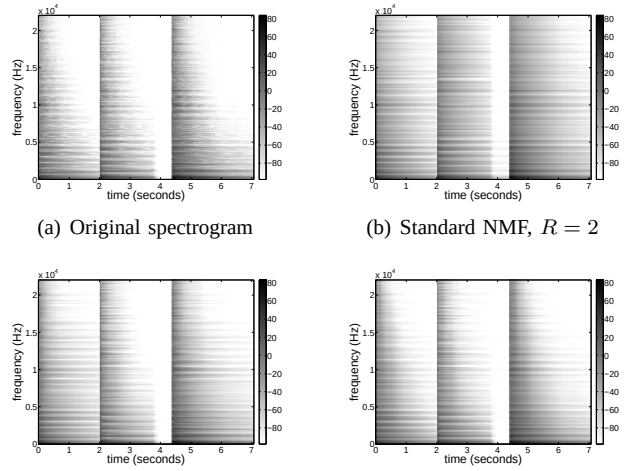


(a) Original spectrogram (b) Standard NMF, $R = 1$

(c) Standard NMF, $R = 5$ (d) Srce/filter factorization, $R = 1$

Fig. 1. Original power spectrogram of the extract of didgeridoo 1(a) and reconstructed spectrograms 1(b), 1(c) and 1(d)

The spectrogram of the extract is represented in figure 2(a). As for most of plucked string instruments, high frequency partials of a harpsichord tone decay faster than low frequency partials. This phenomenon clearly occurs in the L-shaped spectrograms of figure 2(a). The sampling rate of the excerpt is $f_s = 44100Hz$. We chose a 2048-sample-long Hann window with 75% overlap for the STFT.



(a) Original spectrogram (b) Standard NMF, $R = 2$

(c) Standard NMF, $R = 6$ (d) Srce/filter factorization, $R = 2$

Fig. 2. Original power spectrogram of the extract of harpsichord 2(a) and reconstructed spectrograms 2(b), 2(c) and 2(d)

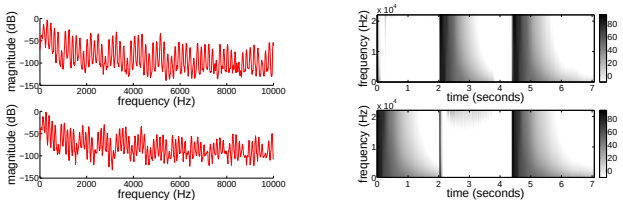
2) Experiment and results: The spectrogram of the excerpt was decomposed with a standard NMF algorithm for $R = 2$ atoms (1 atom per note) and $R = 6$ atoms, and with source/filter factorization for $R = 2$ atoms, with an ARMA modeling ($Q = 1$ and $P = 1$). Reconstructed spectrograms are respectively represented in figures 2(b), 2(c) and 2(d).

The standard NMF needs several atoms per note to accurately decompose the L-shaped power spectrograms: with only 2 atoms (1 per note played), the faster decay of high frequency content does not appear at all (figure 2(b)). With 6 atoms, the attenuation of high frequency partials appears (figure 2(c)),

but each atom is a part of a note spectrum and has no real perceptual meaning.

The ARMA modeling included in our algorithm leads to a good description of the overall spectrogram shape. 2 atoms (1 per note) are enough to accurately fit the original short time spectrum: each atom is harmonic (figure 3(a)) and corresponds to one note while the decay of high frequency partials is clearly well described by the ARMA modeling (see time/frequency activations $h_{rt}^{ARMA}(f)$ in figure 3(b)). The dimensionality of the data provided by our algorithm ($FR+TR(Q+P+1) = 5704$) remains lower than with a standard NMF with 6 atoms ($FR+RT = 9804$) and the global error $\mathcal{C}$ between the original and the reconstructed spectrogram is approximately the same as the one obtained with the standard NMF with $R = 6$.

Thus the decomposition provided by source/filter factorization seems to give a more meaningful representation of the given spectrogram than the one obtained with the standard NMF.



(a) Frequency templates (b) Time/frequency activations

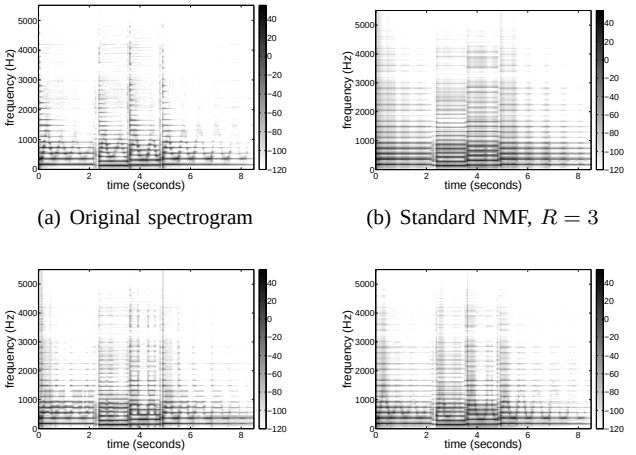
Fig. 3. Source/filter decomposition ($R = 2$, $Q = 1$ and $P = 1$) of the power spectrogram of the harpsichord excerpt

C. Guitar with wah pedal

1) Description of the excerpt: In this section our algorithm is used to decompose a short extract of electric guitar processed by a wah pedal. The wah pedal (or wah-wah pedal) is a popular guitar effect which consists of a resonant filter, the resonant frequency of which is controlled by means of a foot pedal. This effect is named by emphasizing the resemblance with the human onomatopoeia "Wah". A single note of electric guitar processed by a moving wah pedal presents strong spectral variations and therefore cannot be well represented by a single atom in a standard NMF.

As a wah pedal is well modeled by an AR filter with two complex conjugates poles, we chose to decompose the extract with $Q = 0$ and $P = 2$. The chosen extract represented in figure 4(a) is composed of three different notes played successively (the first note is played a second time at the end of the extract). Each note can be viewed as a harmonic pattern which is filtered by a resonant filter, the resonant frequency of which varies between $400Hz$ and $1200Hz$: this resonance clearly appears in the power spectrogram. The sampling rate of the excerpt is $f_s = 11025Hz$. We chose a 1024-sample-long Hann window with 75% overlap for the STFT.

2) Experiment and results: As the analyzed sound presents strong spectral variations, the standard NMF needs many atoms to accurately decompose the power spectrogram. Thus one atom no longer corresponds to one note, and the decomposition does not correspond to the analysis that could



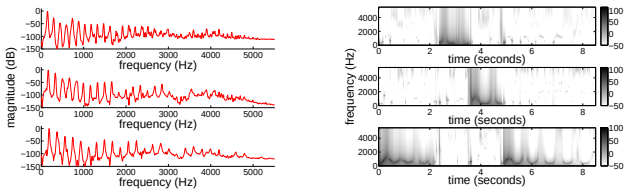
(a) Original spectrogram (b) Standard NMF, $R = 3$

(c) Standard NMF, $R = 10$ (d) Srce/filter factorization, $R = 3$

Fig. 4. Original power spectrogram of the extract of electric guitar processed by a wah pedal 4(a) and reconstructed spectrograms 4(b), 4(c) and 4(d)

be performed by a human listener. Figure 4(b) represents the power spectrogram reconstructed from the NMF of the original spectrogram with 3 atoms and figure 4(c) with 10 atoms. With 3 atoms, NMF is not able to track the resonance of the wah pedal. With 10 atoms, the resonance appears, but the signal is not well explained (each atom is a part of a note and then has no perceptual meaning) and the dimensionality is higher ($FR + RT = 8790$).

With source/filter factorization, the strong spectral variations of each note can be accurately represented in the filter activations taking an order 2 AR modeling ($Q = 0$ and $P = 2$) as shown in figure 4(d). Then 3 atoms (one for each note) are enough to correctly fit the original spectrogram. Indeed, the global β-divergence error between the original and the reconstructed spectrogram obtained with source/filter factorization is approximately the same as the one obtained with standard NMF with 10 atoms; this β-divergence (obtained with source/filter factorization) is also approximately half that obtained with standard NMF with 3 atoms. Each atom is harmonic and corresponds to one note, and the resonance of the wah pedal clearly appears. The dimensionality of the representation obtained with source/filter factorization remains about half that of NMF with 10 atoms: $MR+R(Q+1)N = 4833$. The decomposition provided by our algorithm distinguishes a stationary spectrum representing "average" guitar sounds (contained in $\mathbf{W}$) from the non-stationary effect due to the wah pedal (described by time/frequency activations). The 3 frequency templates (columns of $\mathbf{W}$) obtained are represented in figure 5(a): each template is harmonic with its own fundamental frequency, thus it corresponds to a note (standard NMF with 3 atoms provides similar templates). The time/frequency activations ($h_{rt}^{ARMA}(f)$) are represented in figure 5(b): the resonance of the Wah pedal appears clearly where the notes are played. Thus the decomposition provided by our algorithm seems to give a more meaningful representation of the given spectrogram.

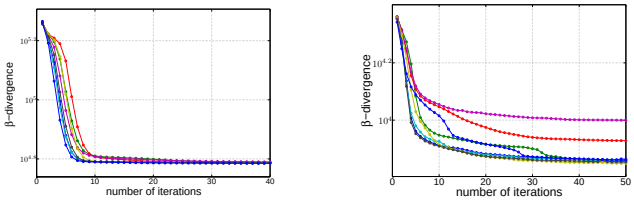


(a) Frequency templates (b) Time/frequency activations

Fig. 5. Source/filter decomposition ($R = 3$ and $P = 2$) of the power spectrogram of the Wah processed guitar excerpt

D. Convergence of the algorithm

The evolution of the cost function over iterations for source filter/factorization is represented in figure 6 with 8 different random initializations, for the decomposition of excerpts presented in sections IV-B (harpsichord excerpt) and IV-C (Wah guitar excerpt). The value of the β-divergence is represented after each iteration. Figures show a monotonic decrease of the cost function and an apparent convergence. In figure 6(a), all initializations lead to the same final value of the cost function and the shape of the evolution is very similar for all initializations. On the other hand, in figure 6(b), all initializations do not lead to the same value of the cost function, showing that multi-initialization is useful.



(a) Harpsichord excerpt (b) Wah guitar excerpt

Fig. 6. Evolution of the cost function over iterations (decomposition of excerpts in sections IV-B and IV-C)

V. Conclusion and Future Work

In this paper, we proposed a new iterative algorithm which is an extended version of the Non-negative Matrix Factorization based on a source/filter representation. We showed that this representation is particularly suitable to efficiently and meaningfully decompose non stationary sound objects including noticeable spectral variations.

In the future, this extended decomposition should be further developed to deal with small pitch variations (like vibrato), for instance using a constant-Q spectrogram like in [23], [24]. Besides, we plan to introduce harmonic constraints in the basis spectra following [25], [26], [27]. Finally, we plan to investigate the introduction of continuity constraints between filters from one frame to another following the approach given in [12], [26], [21].

References

[1] Stéphane Mallat and Zhifeng Zhang, "Matching pursuit with time-frequency dictionnaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[2] Rémi Gribonval and Emmanuel Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 101–111, January 2003.

[3] Pierre Leveau, Emmanuel Vincent, Gaël Richard, and Laurent Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 116–128, January 2008.

[4] P. Comon, "Independent component analysis, a new concept," *Signal Processing special issue Higher-Order Statistics*, vol. 36, pp. 287–314, April 1994.

[5] S. Makino, S. Araki, R. Mukai, and H. Sawada, "Audio source separation based on independent component analysis," in *International Symposium on Circuits and Systems*, Vancouver, Canada, May 2004, pp. V–668–V–671.

[6] Samer A. Abdallah and Mark D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *IEEE Transactions on neural Networks*, vol. 17, no. 1, pp. 179 – 196, January 2006.

[7] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[8] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 2003, pp. 177 – 180.

[9] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorization," in *European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, September 2005.

[10] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, April 2007, vol. 1, pp. I–65 – I–68.

[11] A. Cichocki, R. Zdunek, and S. ichi Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006, vol. 5, pp. 621 – 625.

[12] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

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

[14] Derry FitzGerald, Matt Cranitch, and Eugene Coyle, "Non-negative tensor factorisation for sound source separation," in *Irish Signals and Systems Conference*, 2005.

[15] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Grenada, Spain, September 2004, pp. 494–499.

[16] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 105 – 108.

[17] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, August 2008, pp. 169 – 172.

[18] I. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," in *Neural Information Processing Systems conference (NIPS)*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., pp. 283–290. MIT Press, Cambridge, MA, 2006.

[19] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, March 2007.

[20] A. Cichocki, R. Zdunek, and S. Amari, "Csiszars divergences for non-negative matrix factorization: Family of new algorithms," in *Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Charleston, SC, USA, March 2006, pp. 32 – 39.

[21] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 11, no. 3, pp. 793–830, March 2009.

[22] R. Badeau and B. David, "Weighted maximum likelihood autoregressive and moving average spectrum modeling," in *International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, Nevada, USA, March 2008, pp. 3761 – 3764.

[23] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, March 2008, pp. 2069 – 2072.

[24] M. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Paris, France, April 2006, vol. 3889 of *Lecture Notes in Computer Science (LNCS)*, pp. 700–707, Springer.

[25] Emmanuel Vincent, Nancy Bertin, and Roland Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, March 2008, pp. 109 – 112.

[26] Nancy Bertin, Roland Badeau, and Emmanuel Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[27] Emmanuel Vincent, Nancy Bertin, and Roland Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.