TIME-DEPENDENT PARAMETRIC AND HARMONIC TEMPLATES IN NON-NEGATIVE MATRIX FACTORIZATION

Romain Hennequin, Roland Badeau and Bertrand David, *

Institut Telecom; Telecom ParisTech; CNRS LTCI
Paris France
romain.hennequin@telecom-paristech.fr

ABSTRACT

This paper presents a new method to decompose musical spectrograms derived from Non-negative Matrix Factorization (NMF). This method uses time-varying harmonic templates (atoms) which are parametric: these atoms correspond to musical notes. Templates are synthesized from the values of the parameters which are learnt in an NMF framework. This parameterization permits to accurately model some musical effects (such as vibrato) which are inaccurately modeled by NMF.

1. INTRODUCTION

The decomposition of audio signals in terms of elementary atoms has been a large field of research for years. Sparse decomposition techniques [1] use a redundant dictionary of vectors (called atoms) and try to decompose signals on a few of them (much less than the dimension of the space). When atoms are designed to better encompass some signal properties (for instance harmonic atoms for musical signals [2]), the elements become less generic but more meaningful with regard to the context, and then a supervised classification can be performed to cluster atoms corresponding to a real event in the signal [3].

Recently methods of data factorization were proposed to simultaneously extract atoms from the signal and provide a decomposition on these atoms giving more robustness to the diversity of signals: NMF [4] has been introduced both to reduce the dimensionality and to explain the whole data by a limited number of elementary parts, possibly more significant regarding the considered objects. For instance, thanks to the non-negativity constraint, NMF applied to musical spectrograms will hopefully decompose them into notes, percussive sounds, or rapid transients. This interesting behavior of NMF leads to a wide dissemination of the tool in the audio signal processing community, with a number of applications such as automatic music transcription [5, 6, 7] and sound source separation [8, 9, 10]. Unfortunately, NMF is not well adapted for time-varying phenomena such as vibrato: NMF is a rank reduction technique relying on the frame-to-frame redundancy and slight variations of the fundamental frequency drastically increase the rank by eliminating this redundancy (frequency of high order partials extensively varies and successive frames are then quite dissimilar).

To address this issue, Smaragdis [11] proposes a shift-invariant extension of NMF (in the theoretical background of *Probabilistic Latent Component Analysis*) to decompose constant-Q spectrograms in which transposition can be seen as a shift of the templates. However, the constant-Q spectrogram imposes a quantification of the frequency bins and thus fundamental frequency in a vibrato cannot be estimated with precision. Moreover, the constant-Q transform is not invertible and time-domain signal reconstruction from the decomposition is consequently a difficult task [12].

When used on musical spectrograms, for instance in automatic transcription, "interesting" atoms should hopefully present a specific structure, since musical note spectra possess a harmonic structure. This harmonicity is generally desirable and templates are usually constrained to be harmonic in automatic transcription application [13].

The main idea proposed in this paper is to analytically synthesize such harmonic atoms in a parametric way: this method provides a parametric representation of the harmonic atoms, which can depend on a fundamental frequency parameter, a chirp parameter, a decrease/increase parameter and so on.

The harmonicity constraint is quite strong: the shape of the atoms is restrictive, since they can only be harmonic. However the parameterization should allow for a unique (parametric) atom for a note, the fundamental frequency of which slightly varies (for instance during a vibrato), which was a real issue with standard NMF.

This method provides a representation which can be suitable for automatic transcription, further providing an accurate description of the slight frequency variations.

In section 2, the principle of NMF is briefly reminded and our model is presented as an extension of NMF. In section 3 an algorithm which provides the decomposition is presented. In section 4, the algorithm is used to decompose a musical excerpt showing its ability to deal with slight variations of fundamental frequency (vibrato) and it is then compared to NMF. Finally, conclusions are drawn in section 5.

2. PARAMETRIC TEMPLATE

2.1. Non-negative Matrix Factorization

Given an $F \times T$ non-negative matrix $\mathbf{V}$ and an integer R such that $FR + RT \leq FT$, 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 $\mathbf{H}$:

$$\mathbf{V} \approx \mathbf{WH}, \qquad V_{ft} \approx \hat{V}_{ft} = \sum_{r=1}^{R} w_{fr} h_{rt}. \tag{1}$$

When $\mathbf{V}$ is the magnitude or power spectrogram of a musical signal, the templates that are redundant in multiple frame hopefully are most of the time the harmonic templates corresponding

* The research leading to this paper was supported by the Quaero Programme, funded by OSEO, French State agency for innovation and by the CONTINT program of the French National Research Agency (ANR), as a part of the DReaM project (ANR-09-CORD-006-03).

to musical tones. Thus, each column of $\mathbf{W}$ should correspond to a note and each row of $\mathbf{H}$ is the time activation associated to each note. However, this property is not assured and generally, further constraints are added [13, 14].

2.2. Model

In our model, templates are parameterized following a time-dependent value θ_{rt}. Equation (1) is thus replaced by:

$$V_{ft} \approx \hat{V}_{ft} = \sum_{r=1}^{R} w_{fr}^{\theta_{rt}} h_{rt} \tag{2}$$

where θ_{rt} is the parameter associated to the template r at time t. It can be considered as the "state" of this template: template r will then be synthesized from the value of this parameter for each t. The time-dependence of the parameter now allows for modeling time-varying phenomena such as vibrato.

In this paper, the parameter chosen is the instantaneous fundamental frequency (noted $\theta_{rt} = f_0^{rt}$) of the template: each template is a harmonic comb parameterized by its fundamental frequency.

The parametric template writes:

$$w_{fr}^{f_0^{rt}} = \sum_{k=1}^{n_h(f_0^{rt})} a_k g(f - k f_0^{rt}). \tag{3}$$

This template corresponds in the time domain to a windowed stationary periodic sound (i.e. a windowed sum of sine functions). The Fourier transform of a periodic signal of fundamental frequency f_0 is a sum of Dirac distributions centered in kf_0 (with $k \in \mathbb{Z}^\star$). Thus, when such a signal is windowed, its Fourier transform is the convolution of the previous sum with the Fourier transform of the window. As templates should be non-negative, we take the squared modulus of this Fourier Transform. To make the calculation simpler the modulus of the sum of harmonics is replaced by the sum of the squared modulus of each harmonic: g is the squared modulus of the Fourier transform of the window used in the STFT. The interference between two successive partials is thus neglected; this approximation can be made for sufficiently high fundamental frequencies (or equivalently, sufficiently long analysis windows). The choice of the squared modulus is natural to make function g differentiable, in order to permit standard minimization algorithm.

The expression of function g and its derivative is given in appendix A.1 for a Gauss window and in appendix A.2 for Hann and Hamming windows. Then equation (3) is obtained where a_k are the amplitudes of each harmonic, and $n_h(f_0^{rt})$ the number of harmonics.

Amplitudes a_k of each harmonic are supposed to be the same for every atom and will be learnt in an unsupervised way. It would have been possible to have a different set of partial amplitudes for each template, however, this choice extensively enhances octave (and twelfth, double octave...) ambiguities.

As for a standard NMF, a cost function which is a distance (or a divergence) between the observed spectrogram and the reconstructed spectrogram will be minimized:

$$\mathcal{C}(\mathbf{\Theta}, \mathbf{H}, \mathbf{A}) = D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{ft} d(V_{ft}|\hat{V}_{ft}) \tag{4}$$

where $\mathbf{\Theta} = (\theta_{rt})_{r\in[\![1,R]\!],t\in[\![1,T]\!]}$, $\mathbf{H} = (h_{rt})_{r\in[\![1,R]\!],t\in[\![1,T]\!]}$, and $\mathbf{A} = (a_k)_{k\in[\![1,K]\!]}$.

In this paper, a β-divergence is chosen for d: this is a frequently used class of divergences in NMF frameworks which encompasses a number of usual divergences (Euclidean distance for $\beta = 2$, Kullback-Liebler divergence for $\beta = 1$ and Itakura-Saito divergence for $\beta = 0$). The β-divergence is defined for $\beta \in \mathbb{R}\backslash\{0,1\}$ by:

$$d_\beta(x,y) = \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}).$$

For $\beta \in \{0,1\}$, the value of the β-divergence is the limit of the previous expression.

Then, the cost function is minimized according to h_{rt}, f_0^{rt} and a_k for $r \in [\![1,R]\!]$, $t \in [\![1,T]\!]$ and $k \in [\![1,K]\!]$.

However, the cost function is highly non-convex with respect to (wrt) f_0^{rt}: it can be seen in figure 1 for fixed r and t. We can observe several local minima: two prevailing minima correspond to actually played notes, and several lies near the octaves, the sub-octaves, the twelfths... of each note (This figure can be seen as the opposite of a *spectral product*-like function). Consequently, a global optimization seems to be doomed to failure: a template with a 800Hz initial fundamental frequency will certainly converge to 880Hz. To avoid this issue, rather than a unique generic template for the whole scale, a template by chromatic degree is introduced. Thus, in the proposed decomposition, there is a single harmonic template associated to each chromatic degree. The fundamental frequency of the r^{th} template can vary over time around $f_0^{rt} \approx f_0^{\text{ref}} 2^{\frac{r-1}{12}}$, where f_0^{ref} is the fundamental frequency of the lowest template. Templates with a low activation are discarded afterwards.

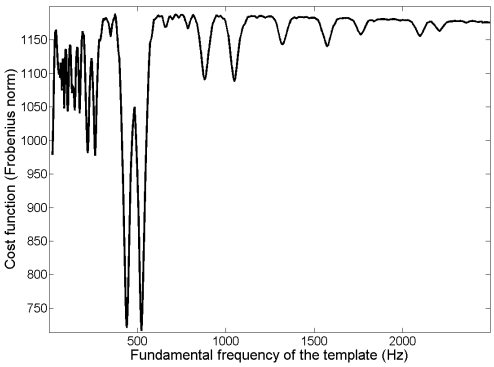

Figure 1: *Cost function wrt the fundamental frequency f_0^{rt} of the templates r: analyzed spectrum is a mix of two harmonic spectra with fundamental frequencies $440Hz$ and $523Hz$*

3. ALGORITHM

Minimization of $\mathcal{C}(\mathbf{\Theta}, \mathbf{H}, \mathbf{A})$ can be done with a multiplicative descent algorithm similar to those generally used for NMF: for the fundamental frequencies, the choice of multiplicative update rules is motivated by the positiveness of the frequency parameter and their natural logarithmic distribution.

In a multiplicative descent algorithm, the update rules associated to one of the parameters λ (here, $\lambda = f(r,t)$, $\lambda = h(r,t)$ or $\lambda = a_k$ for some value of r, t and k) are usually obtained by

expressing the partial derivative of the cost function wrt this parameter as a difference of two positive terms:

$$\frac{\partial \mathcal{C}}{\partial \lambda} = P_\lambda - M_\lambda. \tag{5}$$

The update rule is then:

$$\lambda \leftarrow \lambda \frac{M_\lambda}{P_\lambda}. \tag{6}$$

This rule particularly ensures that λ remains non-negative and becomes constant if the partial derivative is zero.

The partial derivative of the cost function (4) wrt one of the parameter λ is:

$$\frac{\partial \mathcal{C}}{\partial \lambda} = \sum_{ft} \frac{\partial d(V_{ft}, \hat{V}_{ft})}{\partial \lambda} = \sum_{ft} \frac{\partial \hat{V}_{ft}}{\partial \lambda} \frac{\partial d}{\partial y}(V_{ft}, \hat{V}_{ft})$$

where $\frac{\partial d}{\partial y}$ stands for the partial derivative of d wrt to its second argument. With a β-divergence cost function, this partial derivative is:

$$\frac{\partial d}{\partial y}(V_{ft}, \hat{V}_{ft}) = \hat{V}_{ft}^{\beta-2}(\hat{V}_{ft} - V_{ft}).$$

Thus:

$$\frac{\partial \mathcal{C}}{\partial \lambda} = \sum_{ft} \frac{\partial \hat{V}_{ft}}{\partial \lambda} \hat{V}_{ft}^{\beta-2}(\hat{V}_{ft} - V_{ft}).$$

3.1. Update of f_0

In order to obtain the update rules for the parameter $\mathbf{\Theta}$, one needs the partial derivative of the cost function wrt the parameter $\theta_{r_0 t_0}$:

$$\frac{\partial \mathcal{C}}{\partial \theta_{r_0 t_0}} = \sum_{ft} \frac{\partial \hat{V}_{ft}}{\partial \theta_{r_0 t_0}} \hat{V}_{ft}^{\beta-2}(\hat{V}_{ft} - V_{ft}).$$

The partial derivative of the parametric spectrogram $\hat{V}_{ft}$ (see equation (2)) wrt $\theta_{r_0 t_0}$ is given by:

$$\frac{\partial \hat{V}_{ft}}{\partial \theta_{r_0 t_0}} = \delta_{t t_0} h_{r_0 t_0} \frac{\partial w_{f r_0}^{\theta_{r_0 t_0}}}{\partial \theta_{r_0 t_0}}$$

where δ is the Kronecker delta.

When the parameter $\mathbf{\Theta}$ is the fundamental frequency of each template at each time, the partial derivative of the template wrt this fundamental frequency is obtained from (3):

$$\frac{\partial w_{f r_0}}{\partial {f_0}^{r_0 t_0}} = -\sum_{k=1}^{n_h} a_k k g'(f - k {f_0}^{r_0 t_0}).$$

The partial derivative of the cost function wrt ${f_0}^{r_0 t_0}$ is then:

$$\frac{\partial \mathcal{C}}{\partial f_0^{r_0 t_0}} = -\sum_f \sum_{k=1}^{n_h} h_{r_0 t_0} a_k k g'(f - k {f_0}^{r_0 t_0})(\hat{V}_{f t_0}^{\beta-1} - \hat{V}_{f t_0}^{\beta-2} V_{f t_0}).$$

When g has a single lobe (when the window used in STFT is a Gauss window), the following remark is not an issue. However, generally, g has several lobes, and g' changes sign in numerous points. Then, in order to make the calculation easier, only the support of the main lobe of g (noted Λ) is kept in the expression of the derivative:

$$g'(f) \approx g'_a(f) = g'(f) \mathbb{1}_\Lambda(t). \tag{7}$$

Assuming that the main lobe has a unique local maximum (which is actually true for Hamming and Hann windows, see figure 2), function $f \mapsto -g'_a(f)/f$ is then non-negative (see figure 2) and one can write $g'_a(f - k{f_0}^{r_0 t_0}) = -(f - k{f_0}^{r_0 t_0}) P(f - k{f_0}^{r_0 t_0})$, where P is a positive function.

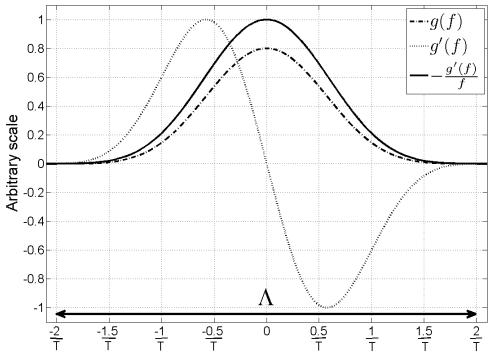

Figure 2: *Main lobe of g, derivative of g and positivity of* $P(f) = -\frac{g'(f)}{f}$ *on* $\Lambda = [-\frac{2}{T}, \frac{2}{T}]$ *for a T-long Hamming window.*[2]

Using approximation (7), the partial derivative of the cost function wrt $f_0^{r_0 t_0}$ is expressible as a difference of two non-negative terms:

$$\frac{\partial \mathcal{C}}{\partial f_0^{r_0 t_0}} \approx \mathcal{G}_{r_0 t_0} - \mathcal{F}_{r_0 t_0} \tag{8}$$

with:

$$\mathcal{G}_{r_0 t_0} = \sum_{f,k} h_{r_0 t_0} a_k k P(f - k {f_0}^{r_0 t_0}) \hat{V}_{f t_0}^{\beta-2}(f \hat{V}_{f t_0} + k {f_0}^{r_0 t_0} V_{f t_0}),$$

$$\mathcal{F}_{r_0 t_0} = \sum_{f,k} h_{r_0 t_0} a_k k P(f - k {f_0}^{r_0 t_0}) \hat{V}_{f t_0}^{\beta-2}(k {f_0}^{r_0 t_0} \hat{V}_{f t_0} + f V_{f t_0}).$$

This results in the update rule of f_0:

$$f_0^{r_0 t_0} \leftarrow f_0^{r_0 t_0} \frac{\mathcal{F}_{r_0 t_0}}{\mathcal{G}_{r_0 t_0}}. \tag{9}$$

The spectrogram will then be decomposed using a single template per semitone. However the update rule of f_0 does not ensure that for each template r, f_0^{rt} remains close to its original fundamental frequency and does not slip to the fundamental frequency band of another semitone. Thus, evolution of f_0^{rt} should be restricted to a fundamental frequency band around the corresponding semitone. In our algorithm, when the fundamental frequency leaves its allocated frequency band, we consider that the corresponding template should not be active at this time and set the corresponding activation to 0.

[2] In order to represent all curves on the same plot, the y-axis scale has been modified and is then arbitrary.

3.2. Update of H

Update rules of $\mathbf{H}$ are obtained in a way similar to standard NMF by computing the derivative of the cost function wrt $h_{r_0t_0}$:

$$\frac{\partial \mathcal{C}}{\partial h_{r_0t_0}} = \sum_{ft} \frac{\partial \hat{V}_{ft}}{\partial h_{r_0t_0}} \hat{V}_{ft}^{\beta-2}(\hat{V}_{ft} - V_{ft}) \quad (10)$$

$$= \sum_{f} w_{fr_0}^{\theta_{r_0t_0}} \hat{V}_{ft_0}^{\beta-2}(\hat{V}_{ft_0} - V_{ft_0}) \quad (11)$$

$$= \mathcal{P}_{r_0t_0} - \mathcal{M}_{r_0t_0} \quad (12)$$

where both $\mathcal{P}_{r_0t_0} = \sum_f w_{fr_0}^{\theta_{r_0t_0}} \hat{V}_{ft_0}^{\beta-1}$ and $\mathcal{M}_{r_0t_0} = \sum_f w_{fr_0}^{\theta_{r_0t_0}} \hat{V}_{ft_0}^{\beta-2} V_{ft_0}$ are positive terms.

Then, the update rule of $h_{r_0t_0}$ is:

$$h_{r_0t_0} \leftarrow h_{r_0t_0} \frac{\mathcal{M}_{r_0t_0}}{\mathcal{P}_{r_0t_0}}. \quad (13)$$

3.3. Update of A

Update rules of $\mathbf{A}$ are obtained in a way similar to the previous section by computing the partial derivative of the cost function wrt a_k:

$$\frac{\partial \mathcal{C}}{\partial a_k} = \sum_{ft} \frac{\partial \hat{V}_{ft}}{\partial a_k} \hat{V}_{ft}^{\beta-2}(\hat{V}_{ft} - V_{ft}). \quad (14)$$

The partial derivative of $\hat{V}_{ft}$ wrt a_k is:

$$\frac{\partial \hat{V}_{ft}}{\partial a_k} = \sum_{r=1}^{R} g(f - kf_0^{rt}) h_{rt} \mathbb{1}_{[1, n_h(f_0^{rt})]}(k).$$

Noting r_k the maximum value of r for which $k \in [1, n_h(f_0^{rt})]$, the previous equation becomes:

$$\frac{\partial \hat{V}_{ft}}{\partial a_k} = \sum_{r=1}^{r_k} g(f - kf_0^{rt}) h_{rt}.$$

Thus, the partial derivative of the cost function wrt a_k can be naturally expressed as a difference of two positive terms:

$$\frac{\partial \mathcal{C}}{\partial a_k} = \mathcal{P}_k - \mathcal{M}_k \quad (15)$$

with:

$$\mathcal{P}_k = \sum_{ft} \sum_{r=1}^{r_k} g(f - kf_0^{rt}) h_{rt} \hat{V}_{ft}^{\beta-1},$$

$$\mathcal{M}_k = \sum_{ft} \sum_{r=1}^{r_k} g(f - kf_0^{rt}) h_{rt} \hat{V}_{ft}^{\beta-2} V_{ft}.$$

The update rule of a_k is then:

$$a_k \leftarrow a_k \frac{\mathcal{M}_k}{\mathcal{P}_k}. \quad (16)$$

3.4. Standard NMF templates to model non-harmonic events

In a musical spectrogram, percussive events (generated by percussive instruments or by the onset of harmonic instruments) do not correspond to harmonic templates and are thus inaccurately taken into account in our model. To encompass this kind of event a standard NMF decomposition term can be added to the parametric spectrogram proposed in (2):

$$V_{ft} \approx \hat{V}_{ft} = \sum_{r=1}^{R} w_{fr}^{\theta_{rt}} h_{rt} + \sum_{r=1}^{R'} w'_{fr} h'_{rt}. \quad (17)$$

Thus w'_{fr} is not time-varying and should model percussive events. R' should be kept very low (in the examples of section 4, $R' = 1$) in order to avoid non-parametric templates representing harmonic events. Update rules of $\mathbf{W}' = (w'_{fr})_{f \in [\![1F]\!] r \in [\![1R']\!]}$ and $\mathbf{H}' = (h'_{rt})_{r \in [\![1R']\!] t \in [\![1T]\!]}$ are standard NMF multiplicative updates for β-divergence which can be found in [15]:

$$\mathbf{W}' \leftarrow \mathbf{W}' . \frac{((\mathbf{W}'\mathbf{H}')^{\cdot\beta-2} . \mathbf{V})\mathbf{H}'^T}{(\mathbf{W}'\mathbf{H}')^{\cdot\beta-1}\mathbf{H}'^T} \quad (18)$$

$$\mathbf{H}' \leftarrow \mathbf{H}' . \frac{\mathbf{W}'^T((\mathbf{W}'\mathbf{H}')^{\cdot\beta-2} . \mathbf{V})}{\mathbf{W}'^T(\mathbf{W}'\mathbf{H}')^{\cdot\beta-1}} \quad (19)$$

where the dot and the fraction bar stands for element-wise operation (element-wise multiplication, element-wise exponent and element-wise division).

3.5. Constraints

As for standard NMF, penalty terms can be added in the cost function to favor certain properties of the decomposition. The update rules are obtained in the same way as presented previously. The partial derivative of the constraint term C_p wrt the parameter λ to be updated is expressed as a difference of two positive terms:

$$\frac{\partial \mathcal{C}_p}{\partial \lambda} = P_\lambda^p - M_\lambda^p.$$

The update rule (6) of the parameter λ thus becomes:

$$\lambda \leftarrow \lambda \frac{M_\lambda + M_\lambda^p}{P_\lambda + P_\lambda^p}$$

where P_λ and M_λ were defined in equation (5).

Several constraints have been considered for the proposed decomposition:

- Sparsity constraint on the columns of H as proposed in [9]: only a few templates should be simultaneously active.
- Uncorrelation constraint between activations of a template and its octave (and eventually twelfth, double octave...) as proposed in [16]: to avoid octave ambiguities, a template should not be active.
- Smoothness constraint on the coefficients of the amplitude as proposed in [9]: the spectral shape of templates should be smooth.

It has been observed that smoothness and uncorrelation constraints noticeably improve the results.

The whole algorithm without penalty term is detailed in Algorithm 1.

Algorithm 1 Time-dependent parametric templates decomposition

Input: $\mathbf{V}$ (spectrogram to be decomposed), R (number of harmonic templates), R' (number of non-harmonic templates), n_{iter} (number of iterations), β (β-divergence)
Output: $\{f_0^{rt}\}_{r\in[\![1,R]\!],t\in[\![1,T]\!]}$, $\mathbf{H}$, $\mathbf{A}$, $\mathbf{W}'$, $\mathbf{H}'$
  Initialize $\mathbf{H}$, $\mathbf{W}'$, $\mathbf{H}'$ with random positive values
  Initialize $\mathbf{A}$ with ones
  Initialize f_0 with normalized frequencies of the chromatic scale: $f_0^{rt} = 2^{\frac{r-1}{12}} f_0^{\text{ref}}$
  for $j = 1$ to n_{iter} **do**
    compute parametric template according to equation (3)
    compute $\hat{\mathbf{V}}$ according to equation (17)
    for all r and t **do**
      update f_0^{rt} according to equation (9)
      if $\left|12\log_2 \frac{f_0^{rt}}{f_0^{\text{ref}}} - (r-1)\right| > 1$ **then**
        set $h_{rt} = 0$
      end if
    end for
    compute parametric template according to equation (3)
    compute $\hat{\mathbf{V}}$ according to equation (17)
    for all k **do**
      update a_k according to equation (16)
    end for
    compute parametric template according to equation (3)
    compute $\hat{\mathbf{V}}$ according to equation (17)
    for all r and t **do**
      update h_{rt} according to equation (13)
    end for
    compute $\hat{\mathbf{V}}$ according to equation (17)
    update $\mathbf{W}'$ with standard NMF rules (equation (18))
    compute $\hat{\mathbf{V}}$ according to equation (17)
    update $\mathbf{H}'$ with standard NMF rules (equation (19))
  end for

4. EXAMPLES

4.1. Decomposition of a musical excerpt

Figure 5 represents the activations h_{rt} of the decomposition of the power spectrogram (represented in figure 3) of an excerpt (four first bars) of the J.S. Bach's first prelude played by a synthesizer. The sampling rate of the excerpt is $f_s = 11025Hz$. We chose a 1024-sample long Hamming window with 75% overlap for the STFT. The decomposition was made with 72 templates (6 octaves) with fundamental frequencies every semitone from 55Hz ($A0$) to 3322Hz ($G\#6$). Kullback-Liebler divergence (β-divergence with $\beta = 1$) was used. Two constraint terms were added: a uncorrelation constraint on activation of octave (and twelfths) and a smoothness constraint on A. All notes were played by the synthesizer with a slight vibrato.

Notes of the piece clearly appear with strong activation values. The maximum simultaneous polyphony is of 3 notes. Even when one note is played simultaneously with its octave, the activation values correspond to effectively played notes. At onset instant, numerous templates are active because no template is able to accurately represent an onset. To reduce this problem, one standard NMF template (non-parametric template) was added to the model in order to better represent the onsets (see section 3.4).

The reconstructed spectrogram is represented in figure 4.

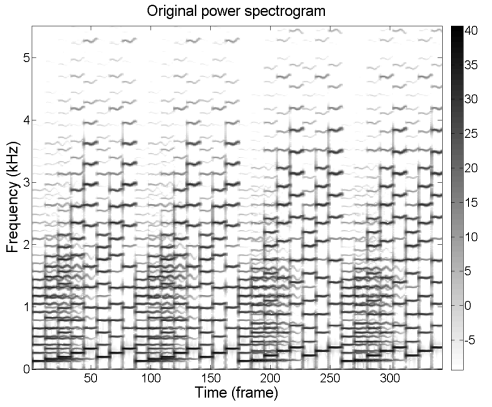

Figure 3: *Original spectrogram of the excerpt of J.S. Bach's first prelude*

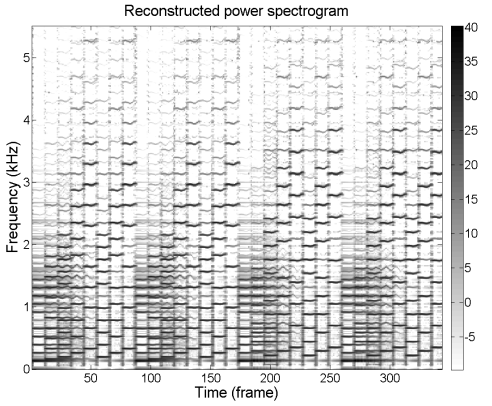

Figure 4: *Reconstructed spectrogram of the excerpt of J.S. Bach's first prelude*

One can easily build a synthetic time/frequency representation of the activations that includes the varying fundamental frequency: for each template r, at each time t, a thin peak is generated in the time/frequency plane at (t, f_0^{rt}) with amplitude equal to the activation h_{rt}. Such a representation is given in figure 6 for the same excerpt in figure 5. This representation reveals the vibrato generated by the synthesizer.

4.2. Comparison with NMF

In this section we will compare our method with the method of decomposition proposed in [13] which is based on NMF with templates that are imposed to be harmonic: in this method, templates are linear combinations of narrow band harmonic patterns. This method is used for transcription and decomposes Equivalent Rectangular Bandwidth (ERB) power spectrograms on 88 templates corresponding each one to a note in the chromatic scale. Smoothness constraints on the activations are added. The input data is not exactly the same (ERB spectrogram in [13] and Standard STFT spectrogram in our method) but the original time-domain signal used is the same and the kind of decomposition provided is very similar in both cases (harmonic templates corresponding to notes

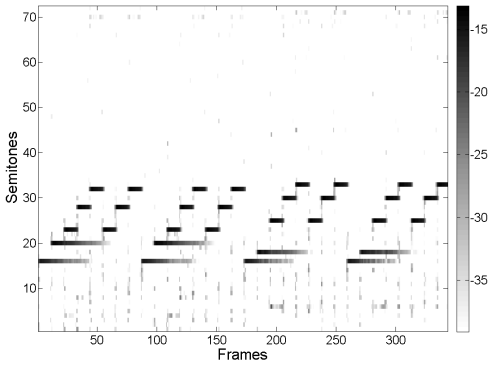

Figure 5: *Activations in the decomposition of the spectrogram of the excerpt of J.S Bach's first prelude. Color scale is in dB. Semitones correspond to relative MIDI note (relative to the lower pitch).*

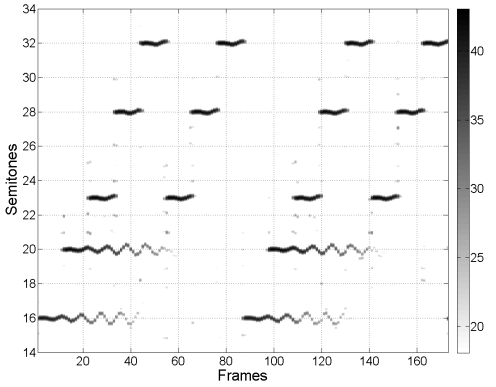

Figure 6: *Representation of the activations including fundamental frequencies (Two first bars of the excerpt of J.S. Bach's first prelude). Color scale is in dB.*

and activation for each template).

The signal analyzed is a C diatonic scale played from $C1$ to $B4$ harmonized with the corresponding third in the scale (each note of the scale is simultaneously played with its third). The scale was played by a synthesizer.

The activation provided by our algorithm is given in figure 7: all notes played clearly appears. The amplitude shape of each note is clear, with a strong onset and a slight decay. One could notice the simultaneous activation of numerous templates at onset times, which is due to the percussive spectral shape of the spectrogram at these instants.

The activation provided by the algorithm of [13] is given in figure 8 with the same dynamic as the activation of our algorithm: most notes are correctly located. However, there are several octave (or sub-octave, twelfth, sub-twelfth...) errors. Moreover the smoothness constraint does not make possible to see the amplitude shape of each note, the onset being spread and the decay being extended. The odd activation of the second template is probably due to the presence of non-harmonic events which are taken into account by this template.

Thus, our decomposition seems to enhance the representation of this signal.

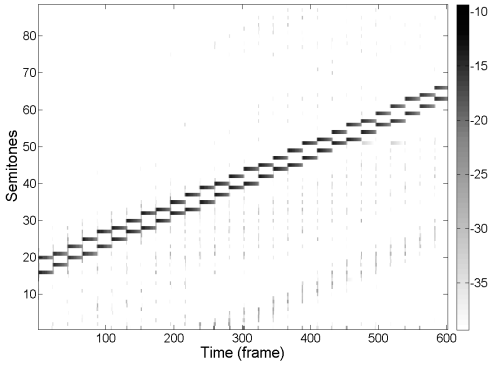

Figure 7: *Representation of the activations provided by our algorithm (amplitude scale is in dB).*

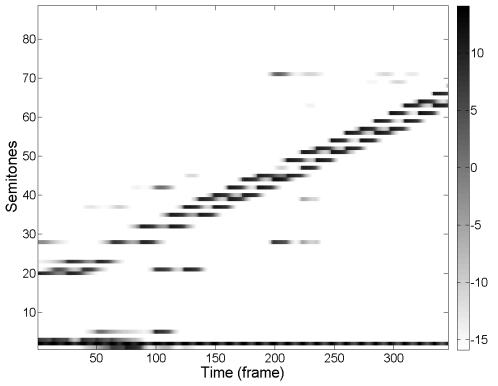

Figure 8: *Representation of the activations provided by the algorithm of [13] (amplitude scale is in dB).*

5. CONCLUSION

In this paper, we presented a new way of decomposing musical spectrograms on a basis of parametric templates which correspond to musical notes. This decomposition provides a good representation of the different notes which are played in each column of the spectrogram. The decomposition being parametric is thus very flexible.

In future work, a supervised learning of atoms could be included in order to improve results and to decompose spectrograms containing several instruments. Moreover, the parameterization of templates could be extended in order to model inharmonicity (that occurs in piano notes), fast fundamental frequency variations (chirp) or fast decrease/increase of the amplitude of the harmonics, and thus to better fit to real audio spectrograms.

6. REFERENCES

[1] Stéphane Mallat and Zhifeng Zhang, "Matching pursuit with time-frequency dictionaries," *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] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[5] Paris Smaragdis and Judith 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.

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

[7] Nancy Bertin, Roland Badeau, and Gaël 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.

[8] Andrzej Cichocki, Rafal Zdunek, and Shun ichi Amari, "New algorithms for nonnegative 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.

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

[10] 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.

[11] Paris Smaragdis, Bhiksha Raj, and Madhusudana 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.

[12] Derry Fitzgerald, Matt Cranitch, and Marcin T. Cychowsky, "Towards an inverse constant q transform," in *Audio Engineering Society Convention Paper*, Paris, France, May 2006.

[13] 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.

[14] 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.

[15] Cédric Févotte, Nancy Bertin, and Jean-Louis 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.

[16] Ye Zhang and Yong Fang, "A NMF algorithm for blind separation of uncorrelated signals," in *International Conference on Wavelet Analysis and Pattern Recognition*, Beijing, China, November 2007, pp. 999–1003.

A. EXPRESSION OF g

In this section, we use the following definition of the Fourier transform of a continuous-time signal x:

$$\hat{X}(f) = \int_{-\infty}^{+\infty} x(t) e^{-i2\pi ft} dt.$$

A.1. Gauss window

In this section, we will give the properties of the Gauss window, *i.e.* the window defined in the time-domain by:

$$h(t) = e^{-\frac{t^2}{\sigma^2}}$$

where σ characterizes the width of the peak.

The Fourier transform of this window is:

$$\hat{H}(f) = \frac{e^{-\sigma^2\pi^2 f^2}}{\sqrt{2}\sigma^2}.$$

Consequently, the expression of g for this type of window is:

$$g(f) = |\hat{H}(f)|^2 = \frac{e^{-2\sigma^2\pi^2 f^2}}{2\sigma^4}.$$

The derivative of g is then:

$$g'(f) = -\frac{2\pi^2 f e^{-2\sigma^2\pi^2 f^2}}{\sigma^2}.$$

For all frequencies $\frac{g'(f)}{f} \leq 0$, which permits to easily write the partial derivative of the cost function wrt the fundamental frequency of a template at a given time as the difference of two positive terms (cf. equation (8)).

The Gauss window has good frequency properties (it has a single main lobe and it saturates the Heisenberg inequality) but are rarely used because of its infinite support and because it does not permit the perfect reconstruction of a signal from its spectrogram.

A.2. "Cosine" window

In this section, we will study the properties of windows defined in the time-domain by:

$$h(t) = (\alpha - \beta\cos(2\pi\frac{t}{T}))\mathbb{1}_{[0,T]}(t)$$

where T is the length of the window, $\mathbb{1}_{[0,T]}(t)$ is the indicator function of the interval $[0, T]$, and $\alpha + \beta = 1$ (the maximum of the window is equal to 1).

This class of windows encompasses:

- Hann window (sometimes referred as Hanning window), for $\alpha = \beta = 0.5$.
- Hamming window, for $\alpha = 0.54$ and $\beta = 0.46$.

The Fourier transform of this window is:

$$\hat{H}(f) = \frac{ie^{-i2\pi Tf}(-1 + e^{i2\pi Tf})(T^2f^2(\beta - \alpha) + \alpha)}{2\pi(T^2f^3 - f)}.$$

Thus, the expression of g for this type of window:

$$g(f) = |\hat{H}(f)|^2 = \frac{1}{4\pi^2}(2-2\cos(2\pi Tf))\frac{(T^2f^2(\beta - \alpha) + \alpha)^2}{f^2(T^2f^2 - 1)^2}.$$

Remark: g can be $\mathcal{C}^1$-prolonged in 0 and $\pm T$ with $g(0) = \alpha^2T^2$ and $g(\pm T) = \frac{\beta^2T^2}{4}$.

The derivative of g is then:

$$g'(f) = \frac{1}{4\pi^2}\left[(2 - 2\cos(2\pi Tf))\frac{2(f^2T^2(\beta - \alpha) + \alpha)}{f^2(f^2T^2 - 1)^2}\right.$$
$$\left(2fT^2(\beta - \alpha) - \frac{2fT^2(f^2T^2(\beta - \alpha) + \alpha)}{f^2T^2 - 1} - \frac{f^2T^2(\beta - \alpha) + \alpha}{f}\right)$$
$$\left. +4\pi T\sin(2\pi Tf)\frac{(T^2f^2(\beta - \alpha) + \alpha)^2}{f^2(T^2f^2 - 1)^2}\right].$$

For Hann and Hamming windows, the main lobe corresponds to frequencies $f \in [-\frac{2}{T}, \frac{2}{T}]$. For these frequencies, $\frac{g'(f)}{f} \leq 0$, which permits to easily write the partial derivative of the cost function wrt the fundamental frequency of a template at a given time as a difference of two positive terms (cf. equation (8)), assuming that g is zero outside the main lobe.