Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription

Nancy Bertin, *Member, IEEE*, Roland Badeau, *Member, IEEE*, and Emmanuel Vincent, *Member, IEEE*

***Abstract*—This paper presents theoretical and experimental results about constrained non-negative matrix factorization (NMF) in a Bayesian framework. A model of superimposed Gaussian components including harmonicity is proposed, while temporal continuity is enforced through an inverse-Gamma Markov chain prior. We then exhibit a space-alternating generalized expectation-maximization (SAGE) algorithm to estimate the parameters. Computational time is reduced by initializing the system with an original variant of multiplicative harmonic NMF, which is described as well. The algorithm is then applied to perform polyphonic piano music transcription. It is compared to other state-of-the-art algorithms, especially NMF-based. Convergence issues are also discussed on a theoretical and experimental point of view. Bayesian NMF with harmonicity and temporal continuity constraints is shown to outperform other standard NMF-based transcription systems, providing a meaningful mid-level representation of the data. However, temporal smoothness has its drawbacks, as far as transients are concerned in particular, and can be detrimental to transcription performance when it is the only constraint used. Possible improvements of the temporal prior are discussed.**

***Index Terms*—Audio source separation, Bayesian regression, music transcription, non-negative matrix factorization (NMF), unsupervised machine learning.**

I. Introduction

NON-NEGATIVE matrix factorization (NMF) is a powerful, unsupervised decomposition technique allowing the representation of two-dimensional non-negative data as a linear combination of meaningful elements in a basis.

NMF has been widely and successfully used to process audio signals, including various tasks such as monaural sound source separation [1], audio stream separation [2], audio-to-score alignment [3], drum transcription [4]. In particular, it has been efficiently used to separate notes in polyphonic music [5], [6] and transcribe it in a symbolic format such as MIDI. In this case, a time–frequency representation of the signal is factored as the product between a basis (or dictionary) of pseudo-spectra and a matrix (decomposition) of time-varying gains. When obtained from harmonic instruments sounds, the basis is shown to partially retain harmonic components, with a pitched structure, that can be interpreted as musical notes, while the decomposition gives information about the onset and offset times of the associated notes.

Meaningful is here a key word: we expect the basis to be formed of interpretable elements, exhibiting certain semantics. The non-negativity constraint is a first step towards this interpretability, compared to other well-known techniques such as singular value decomposition (SVD). For instance, the basis learnt by NMF from an image database is expected to contain meaningful images (the so-called "part-based representation" [7]). This interpretability is often observed in practice, which is certainly one of the reasons for NMF's popularity; but it is not always as satisfying as expected (see, for instance, facial images in [8], that are expected to retain facial parts like eyes, nose, mouth, but do not exactly). As some other desirable characteristics of the decomposition, it is more observed as a welcome side-effect, than enforced and controlled.

To alleviate this lack of control on the decomposition properties, most authors have proposed constrained variants of NMF, ensuring and enhancing those side-effects of baseline NMF: sparsity, spatial localization, temporal continuity, for instance. The typical approach for such constrained variants is to add a penalty term to the usual cost function (reconstruction error) and minimize their sum, see e.g., [1], [8], [9].

On the other hand, several authors have imported the idea of a non-negative constraint in other frameworks than NMF, in particular statistical framework. We can cite non-negative variants of independent component analysis (ICA) [10] and non-negative sparse coding [11]. The Bayesian framework offers both a strong theoretical framework, and the possibility to manage constraints through models and priors.

In this paper, we focus on a Bayesian approach of NMF that allows to enforce harmonicity of the dictionary components (a desired property for music transcription task) and temporal smoothness of the decomposition, preserving however the adaptiveness of NMF, which is purely data-driven, and the interest of the provided mid-level representation for other potential applications. The paper is organized as follows. Section II recalls the baseline NMF model and state-of-the-art constrained NMF algorithms. In particular, constraints of harmonicity and

Manuscript received December 31, 2008; revised September 25, 2009. Current version published February 10, 2010. This work was supported in part by the European Commission under contract FP6–027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content (K-SPACE), and in part by the French GIP ANR under Contract ANR-06-JCJC-0027–01, Décomposition en Éléments Sonores et Applications Musicales (DESAM). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paris Smaragdis.

N. Bertin and R. Badeau are with the Département Traitement du Signal et des Images, Institut TELECOM, TELECOM ParisTech, LTCI CNRS, 75634 Paris, France (e-mail: nancy.bertin@telecom-paristech.fr; roland.badeau@telecom-paristech.fr).

E. Vincent is with INRIA, Centre Inria Rennes Bretagne Atlantique, 35042 Rennes Cedex, France (e-mail: emmanuel.vincent@irisa.fr).

Digital Object Identifier 10.1109/TASL.2010.2041381

temporal continuity are discussed and Bayesian approaches for NMF are presented. Our model, and an EM-like algorithm for NMF with harmonicity and temporal smoothness are proposed in Section III, including a multiplicative initialization phase that updates our previous work on harmonic NMF. Section IV is devoted to experimental results in the transcription task context. The Conclusion and perspectives are drawn in Section V.

II. Constrained Non-Negative Matrix Factorization

Notations

Matrices are denoted by straight bold letters, for instance, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$. Lowercase bold letters denote column vectors, such as $\mathbf{w}_k = (w_{1k} \dots w_{Fk})^T$, while lowercase plain letters with a single index denote rows, such that $\mathbf{H} = (h_1^T \dots h_K^T)^T$. We also define the matrix $\hat{\mathbf{V}} = \mathbf{WH}$.

We use the binary operators $\triangleq$ to denote definitions and $\stackrel{c}{=}$ to denote equality up to an additive constant.

Calligraphic uppercase letters are used to denote probability distributions: $\mathcal{N}$, $\mathcal{P}$, $\mathcal{IG}$ denote Gaussian, Poisson and inverse-Gamma distributions. Their expressions are recalled in Appendix A.

A. Baseline Model and Algorithms

Out of any applicative context, the NMF problem is expressed as follows: given a matrix $\mathbf{V}$ of dimensions $F \times N$ with non-negative entries, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH} = \hat{\mathbf{V}} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $FK + KN \ll FN$, hence reducing the data dimension. In typical audio applications, the matrix $\mathbf{V}$ is chosen as a time–frequency representation (e.g., magnitude or power spectrogram), f denoting the frequency bin and n the time frame.

The factorization (1) is generally obtained by minimizing a cost function defined by

$$D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}) \tag{2}$$

where $d(a|b)$ is a function of two scalar variables. d is typically non-negative and takes value zero if and only if (iff) $a = b$. The most popular cost functions for NMF are the Euclidean (EUC) distance and the generalized Kullback–Leibler (KL) divergence, which were particularly popularized (as NMF itself) by Lee and Seung, see, e.g., [7]. They described multiplicative update rules under which $D(\mathbf{V}|\mathbf{WH})$ is shown to be non-increasing, while ensuring non-negativity of $\mathbf{W}$ and $\mathbf{H}$. The update rules are obtained by using a simple heuristics, which can be seen as a gradient descent algorithm with an appropriate choice of the descent step. By expressing the gradient of the cost function ∇D as the difference of two positive terms $\nabla^+ D$ and $\nabla^- D$, the cost function is shown (in particular cases) or observed to be nonincreasing under the rules:

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla_{\mathbf{W}}^- D(\mathbf{V}|\mathbf{WH})}{\nabla_{\mathbf{W}}^+ D(\mathbf{V}|\mathbf{WH})} \\ \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\nabla_{\mathbf{H}}^- D(\mathbf{V}|\mathbf{WH})}{\nabla_{\mathbf{H}}^+ D(\mathbf{V}|\mathbf{WH})}. \end{cases} \tag{3}$$

For some choices of d, like EUC or KL, monotonicity of the criterion under these rules can be proven [7], but in the general case, these updates do not guarantee any convergence.

B. Constrained Approaches

1) Constraints Imposed via Penalty Terms: In standard NMF, the only constraint is the elementwise non-negativity of all matrices. All other properties of the decomposition, as satisfying as it is, come as uncontrolled side-effects and in a way, the fact that the decomposition retains certain semantics of the original signal, performs separation or provides meaningful and interpretable components is just "good news." It sounds thus natural to try to improve this potential by adding explicit constraints to the factorization problem, in order to enhance and control desired properties.

Then, several constraints have been introduced to get NMF solutions that better fit certain expectancies. Among other proposed constraints, we can cite sparsity [12], spatial localization [8], least correlation between sources [9] or temporal continuity [1], [13].

The common point between those algorithms, whichever constraint is considered, is the "penalty term approach." Rather than minimizing only a reconstruction error term D_r (EUC or KL, typically), the minimized cost function includes a term D_c that quantifies the desired property. The constrained NMF problem is then expressed as

$$\min_{\mathbf{W},\mathbf{H}} \; D_r(\mathbf{V}|\mathbf{WH}) + \lambda D_c(\mathbf{V}|\mathbf{WH})$$

where λ is a weight parameter. Table I gives a few examples of literature penalty terms. Temporal smoothness is one of these examples. In standard NMF and most of its variants, time frames are considered as independent, non-related observations, which is obviously not true for real-world sounds and in particular for music. In the case of musical notes, the main part of the note (the sustain and decay parts, after the attack) possesses a slowly time-varying spectrum. When expressed as the product between a template spectrum $\mathbf{w}_k$ and a time-varying gain h_k, according to NMF formulation, it is equivalent to saying that the row h_k is smooth, or, in other words, that the coefficient h_{kn} is not that different from $h_{k(n-1)}$. [1] and [13] thus introduce penalty terms in the NMF cost function to take into account this temporal continuity. In [1], the term is directly linked to the differences $h_{kn} - h_{k(n-1)}$, while [13] variant relies on a ratio between short-time and long-time variance of h_k. Those terms are shown to favor smoothness in lines of $\mathbf{H}$. Another possible approach is the statistical approach from [14]. Temporal continuity is favored through putting an appropriate prior on $\mathbf{H}$. This

TABLE I
SOME STATE-OF-THE-ART CONSTRAINTS D_c IN NMF PROBLEM

Sparsity	$\sum_{k=1}^{K} \frac{1}{\sqrt{N}-1} \left(\sqrt{N} - \sum_{n=1}^{N} \lvert h_{kn}\rvert / \sqrt{\sum_{n=1}^{N} h_{kn}^2}\right)$	[12]
Spatial localization	$\lambda_1 \sum_{k=1}^{K} \sum_{k'=1}^{K} \left[\mathbf{W}^T\mathbf{W}\right]_{kk'} - \lambda_2 \sum_{k=1}^{K} \left[\mathbf{H}\mathbf{H}^T\right]_{kk}$	[8]
Least correlation	$\sum_{k=1}^{K} \log \left[\mathbf{H}\mathbf{H}^T\right]_{kk} - \log \lvert\mathbf{H}\mathbf{H}^T\rvert$	[9]
Temporal continuity	$\sum_{k=1}^{K} \sum_{n=1}^{N} \lvert h_{kn} - h_{k(n-1)}\rvert^2$	[1]

solution will be exposed with more details and adapted to our case in Section III-C.

It is interesting to notice that non-smoothness may also be an objective (see for instance [15]), depending on the data and the application. [15] points out that smoothness of one of the NMF factors (i.e., $\mathbf{W}$ or $\mathbf{H}$) may enhance sparsity of the other one, thus establishing a link between those two popular constraints. On the other hand, [1] combines sparsity and temporal continuity constraints on $\mathbf{H}$, but concludes to the non-efficiency of the sparsity constraint in his particular case.

The penalty approach has several drawbacks. First, a criterion quantifying the desired property must be found. Second, no general proof of convergence is available for the update scheme (3). Moreover, the parameter λ has to be chosen empirically. These reasons motivated our approach for harmonicity constraint in previous and current work; this approach is exposed in Section II-C2.

2) Deterministic Constraints: Musical notes, excluding transients, are pseudo-periodic. Their spectra are then comb-alike, with regularly spaced frequency peaks. As we wish to use NMF to separate musical notes in a polyphonic recording, we expect that elements in the basis $\mathbf{W}$ are as near as possible from a harmonic distribution. This property is yet not easily quantified by a penalty term.

In [16], we rather proposed an alternative model to baseline NMF problem, enforcing the basis harmonicity. We impose the basis components to be expressed as the linear combination of narrowband harmonic spectra (patterns), which are arbitrarily fixed

$$w_{fk} = \sum_{m=1}^{M} e_{mk} \mathbf{P}_{km}(f). \tag{4}$$

For a given component number k, all the patterns $\mathbf{P}_{km}$ share the same pitch (fundamental frequency f_0); they are defined by summation of the spectra of a few adjacent individual partials at harmonic frequencies of f_0, scaled by the spectral shape of subband k. This spectral envelope is chosen according to perceptual modeling [16]. Fig. 1 illustrates the patterns for one note and the corresponding atom $\mathbf{w}_k$. Coefficients e_{mk} are learned by NMF as well as the decomposition $\mathbf{H}$. Update rules are obtained by minimizing the same cost function as in baseline NMF, except that it is minimized with respect to (wrt) $\mathbf{E}$ and $\mathbf{H}$ rather than $\mathbf{W}$ and $\mathbf{H}$.

3) Statistical Constraints: Another way to induce properties in the NMF is to switch to a statistical framework and introduce adequate prior distributions. Let us consider the following model, proposed in [17], [18]: $\forall n = 1, \ldots, N$

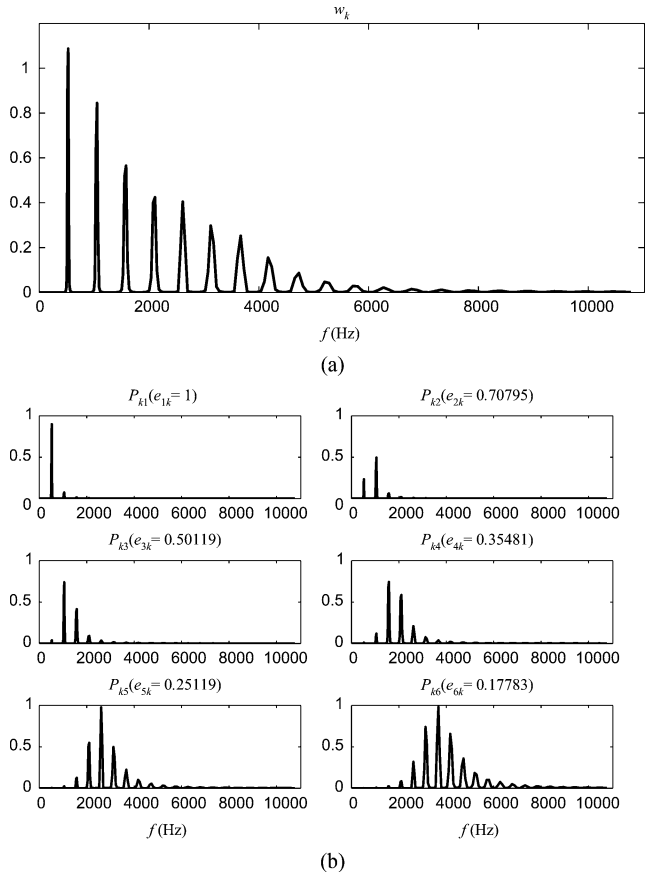

Fig. 1. Example harmonic basis spectrum $\mathbf{w}_k$ corresponding to the note $C4$ (MIDI pitch 72), with underlying narrowband spectra $\mathbf{P}_{km}$ and spectral envelope coefficients e_{mk} (with $M = 6$). (a) Dictionary atom $\mathbf{w}_k$. (b) Corresponding patterns P_{km}.

$$\mathbf{x}_n = \sum_{k=1}^{K} \mathbf{c}_{kn} \in \mathbb{C}^F \tag{5}$$

where latent variables $\mathbf{c}_{kn}$ are independent and follow a multivariate Gaussian distribution

$$\mathbf{c}_{kn} \sim \mathcal{N}(0, h_{kn}\mathrm{diag}(\mathbf{w}_k)). \tag{6}$$

In [14], the estimation of the parameter $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, in a maximum-likelihood (ML) sense is shown to be equivalent to solving the NMF problem $\mathbf{V} \approx \mathbf{WH}$, when observing $\mathbf{V} = (\lvert x_{fn}\rvert^2)_{fn}$ and choosing the underlying cost function d as the Itakura–Saito divergence

$$d_{\mathrm{IS}}(a|b) = \frac{a}{b} - \log\frac{a}{b} - 1. \tag{7}$$

Other authors, like [19], have proven similar equivalences between NMF with KL cost and ML estimation in the model

$$|\mathbf{x}_n| = \sum_{k=1}^{K} |\mathbf{c}_{kn}| \tag{8}$$

under the assumption $|\mathbf{c}_{kn}(f)| \sim \mathcal{P}(w_{fk}h_{kn})$, where $\mathcal{P}(\lambda)$ is the Poisson distribution.

In [20], the authors propose a model where the factors $\mathbf{W}$ and $\mathbf{H}$ are expressed as two functions f_h and f_w (called "link-functions") of Gaussian latent variables. It can be seen as a generalization of the previous model for appropriate choices of f_h and f_w (relatively soft assumptions are put on them). It is another example of the power of the statistical approach to incorporate constraints or knowledge in the NMF problem.

One main advantage of this statistical approach is the possibility to switch from ML estimation to maximum *a posteriori* (MAP) estimation, thanks to Bayes rule

$$p(\mathbf{W},\mathbf{H}|\mathbf{V}) = \frac{p(\mathbf{V}|\mathbf{W},\mathbf{H})p(\mathbf{W})p(\mathbf{H})}{p(\mathbf{V})}. \tag{9}$$

Thus, choosing adequate prior distributions $p(\mathbf{W})$ and $p(\mathbf{H})$ is a way to induce desired properties in the decomposition. Furthermore, the statistical framework provides a strong theoretical basis and efficient algorithms with proven convergence, like the expectation–maximization (EM) algorithm and its variants, to estimate NMF factors.

In next section, we propose to combine this framework and the previous model (4) to enforce both harmonicity in columns of $\mathbf{W}$ and smoothness in rows of $\mathbf{H}$, which are desired properties of the NMF of musical signals.

III. Proposed Algorithm

A. Probabilistic Harmonic Model

The direct usage of formulation (4) in the model (5) is possible, but leads to computational issues. An equivalent model is obtained by assuming

$$\mathbf{x}_n = \sum_{k=1}^{K}\sum_{m=1}^{M} \mathbf{d}_{kmn} \tag{10}$$

with

$$\begin{aligned} \mathbf{x}_n &\in \mathbb{C}^F \\ \mathbf{d}_{kmn} &\sim \mathcal{N}(0, h_{kn}e_{mk}\mathrm{diag}(\mathbf{P}_{km})) \\ \mathbf{P}_{km} &= [P_{km}(1)\dots P_{km}(F)]^T. \end{aligned}$$

Assuming the equality $\mathbf{c}_{kn} = \sum_m \mathbf{d}_{kmn}$ and the independence of $\mathbf{d}_{kmn}$, we can verify that $\mathbf{c}_{kn} \sim \mathcal{N}(0, h_{kn}\sum_m e_{mk}\mathrm{diag}(\mathbf{P}_{km}))$.

From [14], we can establish the equivalence between ML estimation in this generative model (10) and minimization of d_{IS}, which will offer a good coherence and comparability between algorithms for our test. [14] also shows that Itakura–Saito divergence, whose expression is recalled in (7), is well-suited to NMF decomposition of audio signals. Advantages of d_{IS} also include a good fit between the representation and the observation on a log scale (due to the shape of the function $d_{\mathbf{V}}(\hat{v}_{fn}) = d_{\mathrm{IS}}(v_{fn}|\hat{v}_{fn})$ at fixed energy scale v_{fn}, and the strong cost of representing a bin v_{fn} by $\hat{v}_{fn} = 0$) and then, better chances to represent low-energy residual noise (if components are devoted to it, see future work suggestions in Section V). This motivates our model and the choice of IS cost (and not, for instance, the weighted Euclidean cost from [16]) in this work.

B. Maximum-Likelihood Estimation

We now describe an EM-based algorithm for the estimation of the parameters $\boldsymbol{\theta} = \{\mathbf{E},\mathbf{H}\}$. This algorithm is adapted from ML estimation proposed in [14] for the model (10). In ML estimation, the criterion to be maximized is the log-likelihood of the observations

$$C_{\mathrm{ML}}(\boldsymbol{\theta}) \triangleq \log p(\mathbf{V}|\boldsymbol{\theta}). \tag{11}$$

We partition the set of all parameters into disjoint subsets $\boldsymbol{\theta}_k = \{\{e_{mk}\}_m, h_k\}$ so that $\boldsymbol{\theta} = \bigcup_{k=1}^{K}\boldsymbol{\theta}_k$. This partition, and the additive form of the model (10) where the latent variables are supposed independent, allow the usage of the Space Alternating Generalized EM algorithm (SAGE), introduced in [21], to estimate the parameters. The hidden data-space associated with each subset $\boldsymbol{\theta}_k$ is $\mathbf{D}_k = [\mathbf{D}_{k1}\dots\mathbf{D}_{kN}]$, where $\mathbf{D}_{km} = [\mathbf{d}_{km1}\dots\mathbf{d}_{kmN}] \in \mathbb{C}^{F\times N}$. The use of SAGE implies maximizing the functional $Q_k^{\mathrm{ML}}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')$ which is the conditional expectation of the log likelihood of $\mathbf{D}_k$

$$Q_k^{\mathrm{ML}}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \triangleq \int_{\mathbf{D}_k} \log p(\mathbf{D}_k|\boldsymbol{\theta}_k)p(\mathbf{D}_k|\mathbf{V},\boldsymbol{\theta}')d\mathbf{D}_k \tag{12}$$

where $\boldsymbol{\theta}'$ contains the most up-to-date estimated values of all parameters.

We can however notice that Q_k^{ML} can be expressed as the sum (over m) of auxiliary functionals Q_{km}^{ML} expressed as

$$Q_{km}^{\mathrm{ML}}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta}') \triangleq \int_{\mathbf{D}_{km}} \log p(\mathbf{D}_{km}|\boldsymbol{\theta}_{km})p(\mathbf{D}_{km}|\mathbf{V},\boldsymbol{\theta}')d\mathbf{D}_{km} \tag{13}$$

where we define subsets $\boldsymbol{\theta}_{km} = \{e_{mk}, h_k\}$. The problem reduces to maximizing each $Q_{km}^{\mathrm{ML}}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta}')$ wrt e_{mk}, and the sum $Q_k^{\mathrm{ML}}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')$ wrt h_{kn} iteratively. Maximizing these functionals makes the criterion $C_{\mathrm{ML}}(\boldsymbol{\theta})$ increase, according to [21].

At each iteration and for each k, the functionals Q_{km}^{ML} are computed. The sum of the functionals over m is formed and maximized by computing and zeroing its derivative wrt h_{kn}. The derivative wrt e_{mk} of each functional is computed and zeroed, resulting in an update rule for each e_{mk}. Details of the computations are available in Appendix B. Updates rules can be then expressed as follows:

$$h_{kn}^{(\ell+1)} = h_{kn}^{(\ell)} \times \left(1 + \frac{1}{FM}\sum_f\sum_m \frac{h_{kn}^{(\ell)}e_{mk}^{(\ell)}P_{km}(f)}{\hat{v}_{fn}}\left(\frac{v_{fn}}{\hat{v}_{fn}} - 1\right)\right) \tag{14}$$

$$e_{mk}^{(\ell+1)} = e_{mk}^{(\ell)} \times \left(1 + \frac{1}{FN}\sum_n\sum_f \frac{h_{kn}^{(\ell+1)}e_{mk}^{(\ell)}P_{km}(f)}{\hat{v}_{fn}}\left(\frac{v_{fn}}{\hat{v}_{fn}} - 1\right)\right) \tag{15}$$

where the superscript ℓ denotes the value at iteration ℓ and where $\hat{v}_{fn}$ is the current reconstruction of v_{fn}, i.e., $\hat{v}_{fn} = \sum_{k=1}^{K}\sum_{m=1}^{M} h_{kn}e_{mk}P_{km}(f)$ with the most up-to-date values of the parameter, either (ℓ) or $(\ell+1)$ depending on the most recent available values. In SAGE formalism, we update

separately each row h_k, one after the other, *but*, during this step, all h_{kn} for n from 1 to N are updated simultaneously.[1].

Using SAGE framework guarantees the monotonicity of the criterion $C^{\mathrm{ML}}(\boldsymbol{\theta})$. Moreover, [21] proves the existence of a region of monotone convergence in norm, i.e., $\boldsymbol{\theta}$ converges in norm to a local minimum, provided the algorithm was initialized in an appropriate neighborhood of that minimum.

C. Enforcing Temporal Smoothness

In terms of computational cost, this maximum-likelihood estimation of $\mathbf{E}$ and $\mathbf{H}$ has no practical interest, compared to multiplicative gradient descent update rules: as observed in [14] for a similar case (multiplicative versus SAGE algorithm for standard NMF with Itakura–Saito divergence), it is computationally slower and even more sensitive to local minima than usual multiplicative algorithms. However, it has two main advantages: first, the theoretical framework guarantees convergence to a local minimum; second, it opens the possibility of including priors on the parameters, possibly in a hierarchical fashion, and then constraining NMF solutions in an elegant way.

In [14], this framework is exploited to enforce temporal smoothness over the rows of $\mathbf{H}$. We provide *a priori* information on $\boldsymbol{\theta}$, expressed as a prior distribution $p(\boldsymbol{\theta})$. Thanks to the Bayes rule, recalled in (9), we get a MAP estimator by maximizing the following criterion:

$$C_{\mathrm{MAP}}(\boldsymbol{\theta}) \triangleq \log p(\boldsymbol{\theta}|\mathbf{V}) \tag{16}$$

$$\stackrel{\mathrm{c}}{=} C_{\mathrm{ML}}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}). \tag{17}$$

We choose here to use the Markov chain prior structure proposed in [14]

$$p(h_k) = p(h_{k1}) \prod_{n=2}^{N} p(h_{kn}|h_{k(n-1)}) \tag{18}$$

where $p(h_{kn}|h_{k(n-1)})$ reaches its maximum at $h_{k(n-1)}$, thus favoring a slow variation of h_k in time. We proposed for instance the following choice:

$$p(h_{kn}|h_{k(n-1)}) = \mathcal{IG}(h_{kn}|\alpha_k, (\alpha_k+1)h_{k(n-1)}) \tag{19}$$

where $\mathcal{IG}(u|a,b)$ denotes the inverse-Gamma distribution with shape parameter a and scale parameter b, whose mode is $b/(a+1)$; the initial distribution $p(h_{k1})$ is Jeffrey's non-informative prior (see Appendix C).

Several reasons motivated the choice of this prior. First, non-negativity arises naturally from this modeling. Second, this prior is conjugate with respect to the Gaussian observation model, which brings computational simplicity. Moreover, it seems appropriate to the modeling of temporal envelopes in music signal. First, it favors the smoothness by the appropriate choice of the mode. Second, the asymmetry of the inverse-Gamma distribution around the mode (if $\delta h \geq 0$, $p(h_{kn}+\delta h|h_{kn}) \leq p(h_{kn}-\delta h|h_{kn})$) constraints more smoothness on decrease parts $(h_{k(n+1)} \leq h_{kn})$ than on increase parts $(h_{k(n+1)}) \geq h_{kn})$. Thus, it favors smoothness in silence and sustain/decay parts of the notes, but still allows correct representation of the attacks.

[1]With a more explicit notation, at iteration $(\ell+1)$, the coefficient $h_{kn}^{(\ell+1)}$ is determined using, for all p, coefficients $h_{jp}^{(\ell+1)}$ for $j<k$ and $h_{jp}^{(\ell)}$ for $j \geq k$.

TABLE II
COEFFICIENTS OF THE ORDER 2 POLYNOMIAL TO BE SOLVED IN ORDER TO UPDATE h_{kn} IN BAYESIAN HARMONIC NMF WITH AN INVERSE-GAMMA MARKOV CHAIN PRIOR. h_{kn} DENOTES THE ML UPDATE, GIVEN BY THE RIGHT MEMBER OF (14)

	$n=1$	$n=2\dots N-1$	$n=N$
p_0	$\tilde{h}_{k1}$	$\tilde{h}_{kn} + \frac{\alpha_k+1}{FM} h_{k(n-1)}$	$\tilde{h}_{kN} + \frac{(\alpha_k+1)}{FM} h_{k(N-1)}$
p_1	$1 + \frac{1-\alpha_k}{FM}$	$1 + \frac{1}{FM}$	$1 + \frac{1+\alpha_k}{FM}$
p_2	$\frac{1}{FM}\frac{\alpha_k+1}{h_{k2}}$	$\frac{1}{FM}\frac{\alpha_k+1}{h_{k(n+1)}}$	0

Parameters α_k are here arbitrarily fixed, depending on the desired degree of smoothness (the higher α_k, the smoother h_k), but we could consider in future work the possibility to learn it as well.

As the prior respects the scheme $p(\mathbf{H}) = \prod_{k=1}^{K} p(h_k)$, we can still use the SAGE formalism. The functional (12) to minimize is now written

$$Q_k^{\mathrm{MAP}}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \stackrel{\mathrm{c}}{=} \sum_{m=1}^{M} Q_{km}^{\mathrm{ML}}(e_{mk}, h_k|\boldsymbol{\theta}') + \log p(h_k) \tag{20}$$

Q_{km}^{ML} being unchanged, we just have to incorporate the contribution of the prior in the computation and zeroing of the gradients. In Appendix C, this is shown to be proportional to a second-order polynomial

$$\nabla_{h_{kn}} Q_k^{\mathrm{MAP}}(e_{mk}, h_k|\boldsymbol{\theta}') = \frac{-FM}{h_{kn}^2}(p_2\, h_{kn}^2 + p_1\, h_{kn} - p_0). \tag{21}$$

The values of p_0, p_1, p_2 are common for each $n \in [2\dots N-1]$ and take different values at the borders of the Markov chain ($n=1$ and $n=N$). They obviously depend on k, n, and ℓ (though the notation does not mention it, for readability purpose). Their expressions are given in Table II and the detailed computations are available in Appendix C. The resulting update rule is given by the only non-negative root of the polynomial

$$h_{kn}^{(\ell+1)} = \frac{2p_0}{\sqrt{p_1^2 + 4p_2p_0} + p_1} \tag{22}$$

(written here in a form avoiding possible division by zero) and the ML update of $\mathbf{E}$ (15) is unchanged.

In the following, we refer to this algorithm as "Harmonic Smooth NMF" (or, in short form, "HS-NMF").

We can also consider the current model of temporal smoothness, but without harmonicity constraint, leading to the regularized NMF algorithm proposed in [14]. In the following, this algorithm will be denoted as "S-NMF."

D. Multiplicative Initialization With Harmonicity

Due to the slow convergence of EM-like algorithms, HS-NMF needs to be efficiently initialized. Theoretical results from [21] also suggest the interest of smart initialization in terms of convergence of the algorithm. Harmonic multiplicative NMF could then be used to "bootstrap" SAGE algorithm. However, the multiplicative algorithm of [22] was originally designed for a perceptually weighted Euclidean distance, which would not be coherent with HS-NMF criterion (based on IS divergence (7)). For this reason, we wish to adapt harmonic

NMF with multiplicative update rules from [22] to this distance. The criterion to be minimized is written

$$D_{\mathrm{IS}}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_{\mathrm{IS}}\left(v_{fn}|\sum_{k=1}^{K}\sum_{m=1}^{M} h_{kn}e_{mk}P_{km}(f)\right). \quad (23)$$

We compute its derivative wrt h_{kn}, which is expressed as the difference of two positive terms

$$\nabla_{h_{kn}} D_{\mathrm{IS}}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\frac{w_{fk}}{\hat{v}_{fn}} - \sum_{f=1}^{F}\frac{v_{fn}w_{fk}}{\hat{v}_{fn}^2} \quad (24)$$

where $\hat{v}_{fn} = \sum_{k'=1}^{K} w_{fk'}h_{k'n} = \sum_{k'=1}^{K}\sum_{m'=1}^{M} e_{m'k'}P_{k'm'}(f)h_{k'n}$. The derivative wrt e_{mk} fits in the same scheme

$$\nabla_{e_{mk}} D_{\mathrm{IS}}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N}\frac{h_{kn}P_{km}(f)}{\hat{v}_{fn}} - \sum_{f=1}^{F}\sum_{n=1}^{N}\frac{v_{fn}h_{kn}P_{km}(f)}{\hat{v}_{fn}^2}. \quad (25)$$

The update rules are derived from the heuristics (3) and write

$$h_{kn} \leftarrow h_{kn} \times \frac{\sum_{f=1}^{F}\frac{v_{fn}w_{fk}}{\hat{v}_{fn}^2}}{\sum_{f=1}^{F}\frac{w_{fk}}{\hat{v}_{fn}}} \quad (26)$$

$$e_{mk} \leftarrow e_{mk} \times \frac{\sum_{f=1}^{F}\sum_{n=1}^{N}\frac{v_{fn}h_{kn}P_{km}(f)}{\hat{v}_{fn}^2}}{\sum_{f=1}^{F}\sum_{n=1}^{N}\frac{h_{kn}P_{km}(f)}{\hat{v}_{fn}}}. \quad (27)$$

In the following, this algorithm will be referred to as "H-NMF/MU."

IV. Application to Music Transcription

Music transcription consists in converting a raw music signal into a symbolic representation of the music within: for instance a score, or a MIDI file. Here, we focus on information strictly related to musical notes, i.e., musical pitch, onset and offset time, discarding high-level information usually available in a full music sheet, such as bar lines or key signature. Automatic transcription is a very active field of research, known to be difficult, in particular because of note overlapping in the time–frequency plane. Various methods have been proposed to address the transcription issue, including neural network modeling [23], parametric signal modeling and HMM tracking [24] or Bayesian approaches [25]. We propose here to assess the efficiency of Bayesian harmonic and smooth NMF for this task.

A. Experimental Setup

1) Database: To evaluate and quantify transcription performance, we need a set of polyphonic music pieces with accurate MIDI references. The two most simple ways to get such data are either to record a MIDI instrument (the acquisition of audio and MIDI being simultaneous), or to synthesize sound from given MIDI files. For the sake of timbre realism and ease of acquisition, the piano is an instrument of choice: very high quality software synthesizers are available on sale, and an acoustic piano can be equipped to play mechanically, and produce a MIDI output, while retaining the timbre of a real instrument. In his thesis [26], Emiya collected such a database. *MAPS* (MIDI-Aligned Piano Sounds) includes isolated notes, random and tonal chords, pieces from the piano repertoire, recordings on an upright DisKlavier and high-quality software synthesis. From this very complete database, we excerpted two subsets to evaluate our algorithms: a synthetic subset, produced by Native Instruments' Akoustik Piano ("Bechstein Bach" preset, from samples recorded on a Bechstein D280 piano), and a real audio subset, recorded at TELECOM ParisTech on a Yamaha Mark III (upright DisKlavier). Each subset is composed of 30 pieces of 30 s each (original pieces from *MAPS* were truncated). The piano was chosen for practical reasons, but it can be stressed that nothing in the method constraints it to be applied only to piano signals.

2) Structure of NMF-Based Transcription: All NMF-based transcription systems used here follow the same workflow:

1) computation of an adapted time–frequency representation of the signal, $\mathbf{V}$;
2) factorization $\mathbf{V} \approx \mathbf{WH}$;
3) attribution of a MIDI pitch to each basis spectrum $\mathbf{w}_k$ (either from original labeling of columns, when the algorithm includes the harmonicity constraint, or by performing a pitch estimation);
4) onset/offset detection applied to each time envelope h_k.

In [22], it is observed that using a nonlinear frequency scale resulted in a representation of smaller size, with better temporal resolution in the higher frequency range, than usual short-time Fourier transform (STFT), while preserving the subsequent transcription performance. We then pass the signal through a filterbank of 257 sinusoidally modulated Hanning windows with frequencies linearly spaced between 5 Hz and 10.8 kHz on the equivalent rectangular bandwidth (ERB) scale. We then split each subband into disjoint 23-ms time frames and compute the power within each frame.

Pitch estimation of basis spectra is superfluous in NMF with harmonicity constraint, since each basis component can be labeled from the beginning with the pitch of the patterns $\mathbf{P}_{km}$ used to initialize it. For NMF without this constraint, pitch identification is performed on each column of $\mathbf{W}$ by the harmonic comb-based technique used in [16].

Note onsets and offsets are determined by a simple threshold-based detection, followed by a minimum-duration pruning, see [16]. The detection threshold is denoted by A_{dB} and expressed in dB under $\mathbf{H}$ maximum.

3) Evaluation: Transcription performance is quantitatively evaluated according to usual information retrieval scores [27]. **Precision rate** $(\mathcal{P})$ is the proportion of correct notes among all transcribed notes (quantifying the number of notes that are transcribed, but should not). **Recall rate** $(\mathcal{R})$ is the proportion of notes from the MIDI reference which are correctly transcribed (thus quantifying the number of notes that should be transcribed, but are not). **F-measure** $(\mathcal{F})$ aggregates the two former criteria

TABLE III
REFERENCE ALGORITHMS

Abbr.	Description	Reference
NMF/MU	Baseline NMF minimizing IS divergence Multiplicative update rules	[14]
S-NMF	SAGE algorithm for NMF With smoothness constraint on $\mathbf{H}$	[14]
Virtanen'07	Multiplicative NMF With temporal continuity constraint Minimizing KL div. plus a penalty term	[1]
Vincent'08	Multiplicative NMF With weighted Euclidean distance and harmonicity constraint	[16]
Marolt'04	Neural network based transcription	[23]

in one unique score and is defined as $\mathcal{F} = 2\mathcal{PR}/(\mathcal{P}+\mathcal{R})$. A transcribed note is considered as correct if its pitch is identical to the ground truth, and its onset time is within 50 ms of the ground truth, according to community standards (see, for instance, the MIREX competition). Note offset detection is also evaluated through the mean overlap ratio ($\mathcal{MOR}$) defined in [28]. For a correctly transcribed note, the overlap ratio o_{note} between the original note and its transcription is the ratio between the length of the intersection and union of their temporal widths

$$o_{\text{note}} = \frac{\min(t_{\text{off}}) - \max(t_{\text{on}})}{\max(t_{\text{off}}) - \min(t_{\text{on}})} \tag{28}$$

where t_{on} and t_{off} are the vectors of onset times (respectively offset times) of the original and corresponding transcribed note. **Mean Overlap Ratio** ($\mathcal{MOR}$) is the mean of overlap ratios for all correct notes.

The original algorithms (H-NMF/MU and HS-NMF) previously proposed are compared to several state-of-the-art algorithms listed in Table III.

H-NMF/MU, HS-NMF, and S-NMF were implemented by the authors for this work. Virtanen'07 and NMF/MU are run from their author's implementation, which they nicely shared, and Marolt'04 is run from the SONIC software, distributed by its author. Vincent'08 is tuned with the optimal parameters determined in [16]. The order K is set to 88 (the number of components, i.e., of columns in $\mathbf{W}$, is naturally taken as the number of keys on a piano) for all NMF-based algorithms. For algorithms with harmonicity constraint, we take one component (fundamental frequency) per pitch. The maximum number of patterns per note is $M = 10$. When a multiplicative initialization is needed (HS-NMF and S-NMF), ten iterations of the associated multiplicative algorithm (H-NMF/MU and NMF/MU, respectively) are performed before switching to the tested algorithm. Note detection thresholds A_{dB} are manually tuned algorithm per algorithm (and reported in Tables IV and V), by maximizing the average F-measure on each dataset. The minimum duration for a transcribed note is fixed to 50 ms.

B. Results

1) Convergence: We monitor the values of C^{MAP} and D_{IS} at each iteration of HS-NMF, in order to evaluate its speed and efficiency of convergence, and to assess the impact of initializing HS-NMF by H-NMF/MU. Then, we compare the evolution of the criteria between "pure" HS-NMF, and HS-NMF preceded by ten iterations of H-NMF/MU, on the same example

TABLE IV
TRANSCRIPTION SCORES ON SYNTHETIC DATA

Algorithm	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MOR}$	A_{dB}
NMF/MU	63.4	56.1	54.9	51.2	-62
Vincent'08	60.7	60.0	58.4	54.8	-32
H-NMF/MU	58.7	59.1	52.4	46.0	-33
S-NMF	62.4	43.3	49.5	50.7	-51
Virtanen'07	55.9	56.4	53.6	52.1	-22
HS-NMF	65.8	64.5	60.7	44.3	-38
Marolt'04	83.5	70.1	75.8	53.5	-

TABLE V
TRANSCRIPTION SCORES ON REAL AUDIO DATA

Algorithm	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MOR}$	A_{dB}
NMF/MU	43.3	43.4	40.8	47.7	-60
Vincent'08	38.7	37.4	36.1	50.0	-30
H-NMF/MU	43.0	42.7	41.3	44.6	-30
S-NMF	46.2	32.0	36.6	45.6	-49
Virtanen'07	34.2	34.8	33.6	47.1	-21
HS-NMF	46.6	45.3	45.0	43.2	-32
Marolt'04	63.7	53.6	58.0	50.0	-

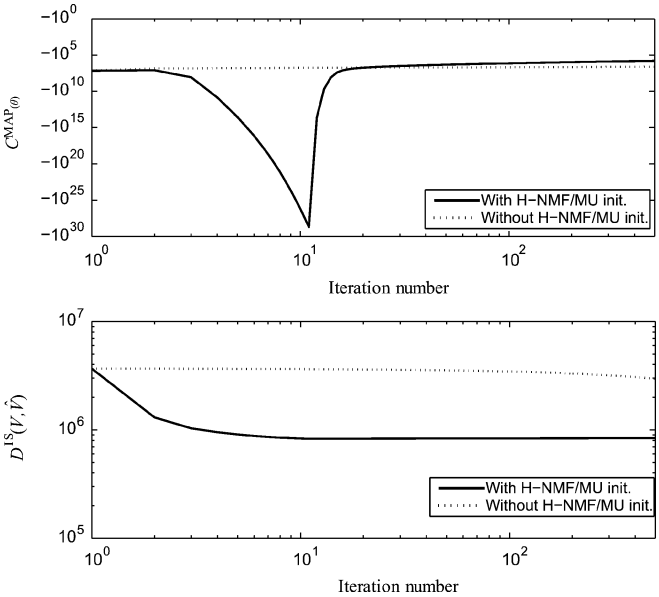

Fig. 2. Evolution of the criteria C^{MAP} and D_{IS} wrt the iteration number.

piece from the dataset and with the same random initialization. Fig. 2 presents this evolution in these two cases. Though C^{MAP} decreases sharply during the initialization (ten first iterations), the multiplicative initialization phase allows the algorithm to reach a higher value of the criterion for the same number of iterations, as well as a lower value of the reconstruction error term D_{IS} (which is equal to the minus log-likelihood up to a constant). After a few hundreds of iterations, the reconstruction error changes very little, while the contribution from the prior still increases slowly, resulting in very few changes in the transcription performance. More decisive, on the presented excerpt (one 30-s piece from the real audio subset), HS-NMF with multiplicative initialization reaches a good transcription performance ($\mathcal{F} = 54.5\%$), while its counterpart without HS-NMF/MU initialization is totally inefficient in separating notes in the same time ($\mathcal{F} = 0\%$ after 500 iterations). An explanation for this is the relative weights between

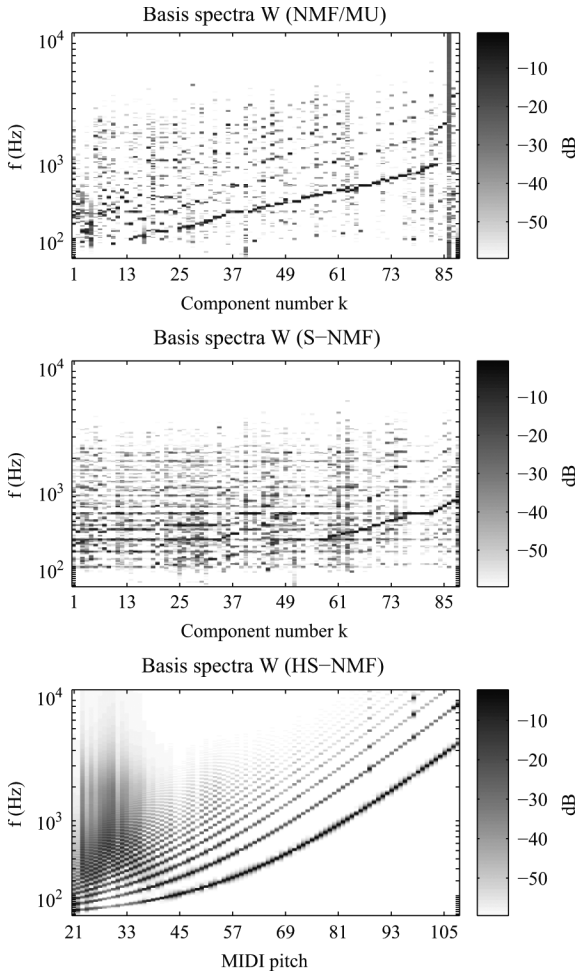

Fig. 3. Example basis matrices $\mathbf{W}$ for algorithms without and with harmonicity constraint. Columns are sorted by increasing pitch.

the two terms in C^{MAP}: the first goal is to reach a good reconstruction, smoothness is a bonus; but if the contribution from the prior takes the most part of the criterion, reconstruction will be poor. Multiplicative initialization allows to optimize first the reconstruction error term, then to focus on the refinement that is the smoothness constraint.

2) Overall Transcription Performance: Tables IV and V report the transcription performance of tested algorithms on the synthetic and recorded datasets, respectively. HS-NMF outperforms other NMF-based algorithms in both cases, but remains less performant than SONIC software. Smoothness constraint used alone seems detrimental to transcription performance, may it be implemented by a multiplicative algorithm (Virtanen'07) or by a Bayesian algorithm (S-NMF), but improves the performance of harmonically constrained NMF (H-NMF versus HS-NMF).

Results are comparable to scores from [24] obtained on a database including ours, and place our algorithm performance at the state-of-the-art level.

3) Harmonicity of the Basis: In Fig. 3, we display bases $\mathbf{W}$ after convergence, with columns sorted by increasing pitch. We can see that non-harmonically constrained NMF exhibits a dictionary that has a pitched structure but a rather noisy look, whereas bases from harmonically constrained algorithms are much cleaner. S-NMF produces a much less sparser dictionary than unconstrained NMF, which is coherent with observations from [15] and could explain its lower performance. Another noticeable result is the pitch repartition in the basis. In NMF without harmonicity constraint, as the basis is completely free, pitch repartition in the basis follows the same trend as pitch repartition in the original piece; NMF tends to use more components to represent faithfully the most frequent notes, while possibly neglecting rare passing tones. Moreover, some components do not exhibit a pitched structure (5, in average). On the contrary, NMF with harmonicity constraint have a fixed number of components per pitch (one, in our case). This guarantees representation of all notes, including notes played only a few times in the piece, but implies also useless computation on components corresponding to absent notes in the piece, and does not allow representation of non-harmonic parts of the signal. This could be alleviated by adding unconstrained components to the harmonic dictionary, updated separately under usual multiplicative rules, for instance.

4) Smoothness of Components: Temporal envelopes h_k, for k corresponding to the note $C4$, obtained by NMF/MU (without constraint), H-NMF, S-NMF, and HS-NMF are displayed in Fig. 4. The ground truth pianoroll (time–pitch representation) is displayed as well. S-NMF and HS-NMF produce indeed smoother envelopes, which can be noticed in particular when the note is supposed to be off. We can notice several spurious peaks in NMF/MU and H-NMF/MU, for instance during the first 750-ms [region (a)] or around $t = 10$ s [region (b)], whose amplitude is reduced or zeroed by the associate smooth version (S-NMF and HS-NMF, respectively). Another noticeable result is that harmonicity constraint seems to disfavor smoothness of the envelopes. We also briefly investigated on the impact of the choice of α_k on smoothness and performance; values of α_k between 5 and 15 resulted in a loss of less than 2 points in the F-measure and a barely noticeable difference in the smoothness of rows.

We could have expected a positive influence of the smoothness constraint on the $\mathcal{MOR}$ values, which would mean a better offset detection. Though we observe slightly better $\mathcal{MOR}$ for Virtanen'07 and S-NMF compared to, for instance, HS-NMF, it seems here difficult to draw a straightforward conclusion, partially because of the previously observed interactions between harmonicity and smoothness constraints.

5) Detection Threshold: In Tables IV and V, the optimal detection threshold is manually determined to get the best mean F-measure over the test database. Varying this threshold allows to display Precision–Recall curves and have a deeper insight on algorithms performance. Fig. 5 presents these curves for NMF/MU, H-NMF/MU, S-NMF, and HS-NMF. The curve confirms the good performance of HS-NMF. It reaches a better tradeoff between precision and recall and is more robust to the choice of the threshold. Both multiplicative algorithms (H-NMF/MU and NMF/MU) are comparable around the optimal F-measure. S-NMF gives the poorest results at every threshold. We can also notice that a 100% recall is never

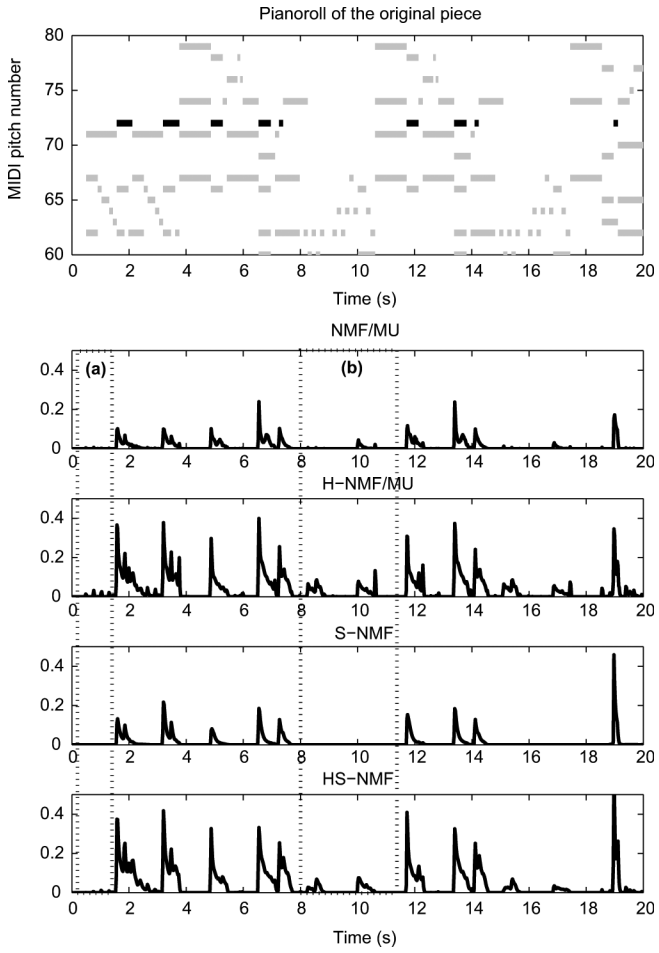

Fig. 4. Temporal activation of note $C4$ for four different algorithms (NMF/MU, H-NMF/MU, S-NMF, and HS-NMF from top to bottom) on the same excerpt. The pianoroll of the corresponding excerpt is on top, with $C4$ in black and neighbor notes in gray. Regions of interest are framed with dotted lines.

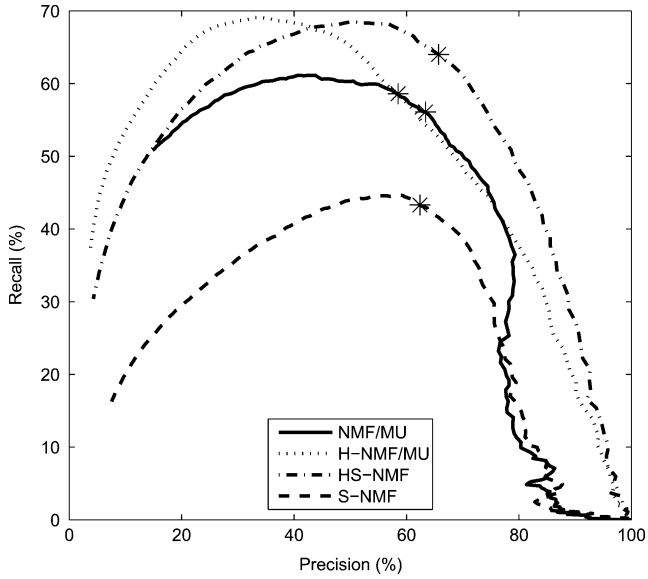

Fig. 5. Precision-Recall curves for four different algorithms. The detection threshold varies from 0 to -100 dB under $\mathbf{H}$ maximum. The couple $(\mathcal{P},\ \mathcal{R})$ realizing the $\mathcal{F}$ maximum is represented with a star.

reached, even at very low threshold, which points a limit of NMF-based transcription algorithms.

These curves, as well as Tables IV and V, are obtained by averaging the scores over the dataset, but it is important to note an important variability between pieces, in terms of performance and optimal threshold. At fixed threshold A_{dB}, $\mathcal{F}$ standard deviation is worth about 12% for all NMF-based algorithms (from 9% for Virtanen'07, to 16% for HS-NMF).

V. Conclusion and Perspectives

In this paper, we proposed an original model for including harmonicity and temporal smoothness constraints in non-negative matrix factorization of time–frequency representations, in a unified framework. The resulting algorithm we propose, HS-NMF, is derived from a Bayesian framework and outperforms other benchmarked NMF approaches in a task of polyphonic music transcription, evaluated on a realistic music database. The Bayesian framework also offers theoretical results about convergence, that are generally not available in usual multiplicative approaches of NMF. We also proposed a novel multiplicative NMF with harmonicity constraint, minimizing Itakura–Saito divergence, which has links with the exposed statistical approach and was shown to suit well for the representation of audio signals in this context [14]. Thus, the contributions of this paper are theoretical, algorithmic and experimental at a time, in the very active domains of music transcription and NMF study.

NMF-based methods remain here less performant than other finely tuned state-of-the-art methods, especially methods implying a training phase, the use of learning data and musicologically inspired postprocessing. However, NMF is totally data-driven, it requires no training and then adapts itself to the data while avoiding the risk of a mismatch between training and test data. It also provides a semantically meaningful mid-level representation of the data. Its potential here assessed is clear, letting the hope of very good performance with better tuning and improvements. The temporal smoothness constraint does not bring all improvements we could expect, in particular in terms of robustness to the detection threshold and efficiency of the note duration estimation. However, it seems useful to compensate the tendency of NMF with harmonicity constraint to produce non-smooth decomposition, and lead therefore to a better transcription performance when both constraints are used. A limitation of our common NMF framework (NMF core algorithm plus detection threshold based postprocessing) appears here, as a 100% recall rate is never reached, for any value of the threshold or any tested algorithm.

Using a statistical model relies of course on the fact that the ground truth actually follows this model. Performance obtained here let hope it is more or less the case, but adequation between the data and the model should be further investigated on. In particular, the choice of the shape parameter α of the inverse-Gamma prior put on temporal envelopes should be discussed, and its learning, as well as NMF factors are learned, should be considered.

Possible improvements include a refinement of the temporal prior, which suits for modeling the sustain and decay parts of

the note, but disfavor attacks and silences. An option to alleviate this mismatch between the model and the data could be the use of switching state models for the rows of $\mathbf{H}$, that would explicitly model the possibility for h_{kn} to vary quickly (attack) or to be strictly zero (absence of the note). As far as $\mathbf{W}$ is concerned, transients are badly represented in an entirely harmonic dictionary, but this could be solved by adding a few unconstrained (non harmonic) components in the representation, which would hopefully be well captured thanks to d_{IS} scale-invariance. At last, as many EM-based algorithms, HS-NMF remains very slow compared to multiplicative gradient descent approaches; an alternative to it could be the direct minimization of the criterion it optimizes by the usual multiplicative heuristics (3), possibly losing the proof of convergence but reducing computational time.

Appendix A
Standard Distributions

Complex valued Gaussian $\mathcal{N}(\mathbf{u}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = |\pi\,\boldsymbol{\Sigma}|^{-1} \exp -(\mathbf{u}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})$
Poisson $\mathcal{P}(u|\lambda) = \exp(-\lambda)\,\lambda^u/u!$
Inverse-Gamma $\mathcal{IG}(u|\alpha,\beta) = \beta^\alpha/\Gamma(\alpha)\,u^{-(\alpha+1)}\,\exp(-\beta/u)$, $u \geq 0$.

Appendix B
SAGE Update Rules With Harmonicity

In this Appendix , we detail the derivations leading to update rules of (14) and (15). The functional $Q_{km}^{\mathrm{ML}}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta'})$ defined in (13) may be processed in two steps. First, we write the hidden data log-likelihood

$$\log p(\mathbf{D}_{km}|\boldsymbol{\theta}_{km}) = \log \prod_{n=1}^{N}\prod_{f=1}^{F} p(d_{kmn}(f)|\boldsymbol{\theta}_{km}). \quad (29)$$

As $d_{kmn}(f) \sim \mathcal{N}(0, h_{kn}e_{mk}\mathbf{P}_{km}(f))$, we have

$$\log p(\mathbf{D}_{km}|\boldsymbol{\theta}_{km}) \stackrel{\mathrm{c}}{=} -\sum_{n=1}^{N}\sum_{f=1}^{F}\log(h_{kn}e_{mk}\mathbf{P}_{km}(f)) + \frac{|d_{kmn}(f)|^2}{h_{kn}e_{mk}\mathbf{P}_{km}(f)}. \quad (30)$$

The second term to be computed is the hidden data posterior $p(\mathbf{D}_{km}|\mathbf{V},\boldsymbol{\theta'})$. It may be obtained by writing $\mathbf{x}_n = \mathbf{d}_{kmn} + \sum\sum_{(k',m')\neq(k,m)} \mathbf{d}_{k'm'n}$ and using the Wiener filtering method proposed in [17] for the separation of two sources. According to it, the posterior mean and variance of $d_{kmn}(f)$ write, respectively,

$$\mu_{kmn}^{\mathrm{post}}(f) = \frac{h_{kn}e_{mk}\mathbf{P}_{km}(f)}{\hat{v}_{fn}}x_n(f) \quad (31)$$

$$\lambda_{kmn}^{\mathrm{post}}(f) = \frac{h_{kn}e_{mk}\mathbf{P}_{km}(f)}{\hat{v}_{fn}} \times \sum_{(k',m')\neq(k,m)}\sum h_{k'n}e_{k'm'}\mathbf{P}_{k'm'}(f). \quad (32)$$

Then, by taking the expectation of the log-likelihood with regard to the posterior, we get the functional expression

$$Q_{km}^{\mathrm{ML}}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta'}) = -\sum_{n=1}^{N}\sum_{f=1}^{F}\log(h_{kn}e_{mk}\mathbf{P}_{km}(f)) + \frac{\left|\mu_{kmn}^{\mathrm{post}}(f)\right|^2 + \lambda_{kmn}^{\mathrm{post}}(f)}{h_{kn}e_{mk}\mathbf{P}_{km}(f)}. \quad (33)$$

Zeroing the gradients of Q_{km}^{ML} wrt e_{mk} and the gradient of their sum over m wrt h_{kn} leads to the update rules

$$h_{kn}^{(\ell+1)} = \frac{1}{FM}\sum_f\sum_m \frac{\left|\mu_{kmn}^{\mathrm{post}'}(f)\right|^2 + \lambda_{kmn}^{\mathrm{post}'}(f)}{e_{mk}^{(\ell)}\mathbf{P}_{km}(f)} \quad (34)$$

$$e_{mk}^{(\ell+1)} = \frac{1}{FN}\sum_n\sum_f \frac{\left|\mu_{kmn}^{\mathrm{post}'}(f)\right|^2 + \lambda_{kmn}^{\mathrm{post}'}(f)}{h_{kn}^{(\ell+1)}\mathbf{P}_{km}(f)} \quad (35)$$

where the superscript $'$ indicates that $\lambda_{kmn}^{\mathrm{post}'}$ and $\mu_{kmn}^{\mathrm{post}'}$ are computed with most up-to-date values of $\mathbf{E}$ and $\mathbf{H}$. This form lets appear possible numeric errors if $h_{kn} = 0$ or $e_{mk} = 0$. This can be avoided by replacing $\lambda_{kmn}^{\mathrm{post}}$ and $\mu_{kmn}^{\mathrm{post}}$ by their expressions (31) and (32). This leads to update rules proposed in (14) and (15).

Appendix C
SAGE Update Rules With Harmonicity and Temporal Smoothness

We write the functional $Q_k^{\mathrm{MAP}} = \sum_{m=1}^{M} Q_{km}^{\mathrm{ML}} + \log p(h_k)$ as the sum of the ML functional and contributions from the prior. For $n = 2 \ldots N-1$

$$\nabla_{h_{kn}} Q_k^{\mathrm{MAP}}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) = \nabla_{h_{kn}}\left(\sum_{m=1}^{M} Q_{km}^{\mathrm{ML}}(\boldsymbol{\theta}_{km}|\boldsymbol{\theta'})\right) + \nabla_{h_{kn}}\left(\log p(h_{k(n+1)}|h_{kn}) + \log p(h_{kn}|h_{k(n-1)})\right). \quad (36)$$

As $\log \mathcal{IG}(u|\alpha,\beta) \stackrel{\mathrm{c}}{=} \alpha\log\beta - (\alpha+1)\log u - \beta/u$, we have (37), as shown at the top of the next page. Then, this gradient is proportional to a second-order polynomial

$$\nabla_{h_{kn}} Q_k^{\mathrm{MAP}}(\boldsymbol{\theta}_k|\boldsymbol{\theta'}) = \frac{-FM}{h_{kn}^2}\left(p_2 h_{kn}^2 + p_1 h_{kn} - p_0\right)$$

$$\text{with} \quad p_2 = \frac{1}{FM}\frac{\alpha_k+1}{h_{k(n+1)}}$$

$$p_1 = 1 + \frac{1}{FM}$$

$$p_0 = \tilde{h}_{kn} + \frac{\alpha_k+1}{FM}h_{k(n-1)}$$

where $\tilde{h}_{kn}$ is the ML estimator (see (34)). For $n = N$ the term $p(h_{k(n+1)}|h_{kn})$ is simply removed from (36). For $n = 1$, the Markov chain structure imposes to choose a prior $p(h_{k1})$. We take Jeffreys' non-informative prior: $p(h_{k1}) \propto 1/h_{k1}$. The corresponding gradients are written, as shown in the equation at the top of the next page.

This leads to p_0, p_1 and p_2 values reported in Table II.

$$\nabla_{h_{kn}} Q_k^{\mathrm{MAP}}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') = -\frac{\alpha_k+1}{h_{k(n+1)}} - \frac{FM+1}{h_{kn}} + \frac{1}{h_{kn}^2} \times \left(\sum_{f=1}^{F} \sum_{m=1}^{M} \frac{\left|\mu_{kmn}^{\mathrm{post}}(f)\right|^2 + \lambda_{kmn}^{\mathrm{post}}(f)}{e_{mk} P_{km}(f)} + (\alpha_k+1) h_{k(n-1)} \right). \quad (37)$$

$$\nabla_{h_{k1}} Q_k^{\mathrm{MAP}}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') = -\frac{FM}{h_{k1}} + \frac{1}{h_{k1}^2} \left(\sum_{f=1}^{F} \sum_{m=1}^{M} \frac{\left|\mu_{kmn}^{\mathrm{post}}(f)\right|^2 + \lambda_{kmn}^{\mathrm{post}}(f)}{e_{mk} \mathbf{P}_{km}(f)} \right) - \frac{\alpha_k-1}{h_{k1}} - \frac{\alpha_k+1}{h_{k2}}.$$

ACKNOWLEDGMENT

The authors would like to thank C. Févotte for his decisive influence on the Bayesian orientation of this work, wise advice on literature review, and support. They would also like to credit V. Emiya for its incredible work on collecting and sharing MAPS database, and T. Virtanen for gently sharing code and usage advice.

REFERENCES

[1] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[2] B. Wang and M. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in *Proc. DMRN Summer Conf.*, Glasgow, U.K., Jul. 23–24, 2005.

[3] A. Cont, "Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, Toulouse, France, May 14–17, 2006, pp. 245–248.

[4] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proc. 13th Eur. Signal Process. Conf. (EUSIPCO)*, Antalya, Turkey, Sep. 4–8, 2005.

[5] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'03)*, New Paltz, NY, Oct. 19–22, 2003, pp. 177–180.

[6] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, Honolulu, HI, Apr. 15–20, 2007, vol. 1, pp. 65–68.

[7] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[8] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition*, Dec. 11–13, 2001, vol. 1, pp. 207–212.

[9] Y. Zhang and Y. Fang, "A NMF algorithm for blind separation of uncorrelated signals," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognition*, Beijing, China, Nov. 2–4, 2007, pp. 999–1003.

[10] M. Plumbley, "Algorithms for nonnegative independent component analysis," *IEEE Trans. Neural Netw.*, vol. 14, no. 3, pp. 534–543, Mar. 2003.

[11] S. Abdallah and M. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proc. 5th Int. Conf. Music Inf. Retrieval (ISMIR'04)*, Barcelona, Spain, Oct. 10–14, 2004, pp. 318–325.

[12] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Nov. 2004.

[13] Z. Chen, A. Cichocki, and T. M. Rutkowski, "Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer's disease," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, Toulouse, France, May 14–19, 2006, vol. 5, pp. 893–896.

[14] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.* [Online]. Available: http://www.tsi.enst.fr/~fevotte/TechRep/techrep08 is-nmf.pdf

[15] A. Pascual-Montano, J. Carazo, K. Kochi, D. Lehmann, and R. Pascual-Marqui, "Nonsmooth nonnegative matrix factorization (nsNMF)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 403–415, Mar. 2006.

[16] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Las Vegas, NV, Apr. 4, 2008, pp. 109–112.

[17] L. Benaroya, L. McDonagh, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, Hong Kong, China, Apr. 6–10, 2003, pp. 613–616.

[18] L. Benaroya, R. Blouet, C. Févotte, and I. Cohen, "Single sensor source separation using multiple-window STFT representation," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC'06)*, Paris, France, Sep. 12–14, 2006.

[19] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to nonnegative matrix factorisation for audio signal modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Las Vegas, NV, Apr. 4, 2008, pp. 1825–1828.

[20] M. Schmidt and H. Laurberg, "Nonnegative matrix factorization with Gaussian process priors," *Comput. Intell. Neurosci.*, 2008.

[21] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation maximization algorithm," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2664–2677, Oct. 1994.

[22] E. Vincent, N. Bertin, and R. Badeau, "Two nonnegative matrix factorization methods for polyphonic pitch transcription," in *Proc. Music Inf. Retrieval Evaluation eXchange (MIREX)*, Austria, Sep. 23–30, 2007.

[23] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, Jun. 2004.

[24] V. Emiya, R. Badeau, and B. David, "Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches," in *Proc. Eur. Conf. Sig. Process. (EUSIPCO)*, Lausanne, Switzerland, Aug. 25–29, 2008.

[25] A. Cemgil, H. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[26] V. Emiya, "Transcription automatique de la musique de piano," Ph.D. dissertation, Institut TELECOM; TELECOM ParisTech, Paris, France, 2008.

[27] C. van Rijsbergen, *Information Retrieval*, 2nd ed. London, U.K.: Butterworths, 1979.

[28] M. Ryynänen and A. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2005, pp. 319–322.

Nancy Bertin (M'06) was born in France on August, 1st 1981. She received the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2004, the M.Sc. degree in acoustics, computer science, and signal processing applied to music (ATIAM) from the Université Pierre et Marie Curie (Paris VI), Paris, in 2005, and the Ph.D. degree in signal processing from Télécom ParisTech (ENST), Paris, in 2009.

Her interests included music information retrieval, blind signal decompositions, and statistical models for music signals. She is now part of the METISS group, INRIA, Rennes, France, where her research focuses on source separation and compressed sensing of acoustic fields.

Roland Badeau (M'02) was born in Marseilles, France, on August 28, 1976. He received the State Engineering degree from the École Polytechnique, Palaiseau, France, in 1999, the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001, the M.Sc. degree in applied mathematics from the École Normale Supérieure (ENS), Cachan, France, in 2001, and the Ph.D. degree from the ENST in 2005, in the field of signal processing.

In 2001, he joined the Department of Signal and Image Processing, TELECOM ParisTech (ENST), as an Assistant Professor, where he became Associate Professor in 2005. His research interests include high-resolution methods, adaptive subspace algorithms, audio signal processing, and music information retrieval.

Emmanuel Vincent (M'07) received the mathematics degree from the École Normale Supérieure, Paris, France, in 2001 and the Ph.D. degree in acoustics, signal processing, and computer science applied to music from the University of Paris-VI Pierre et Marie Curie, Paris, in 2004.

From 2004 to 2006, he was a Research Assistant with the Centre for Digital Music at Queen Mary, University of London, London, U.K. He is now a Permanent Researcher with the French National Institute for Research in Computer Science and Control (INRIA). His research focuses on probabilistic modeling of audio signals applied to blind source separation, indexing, and object coding of musical audio.