Harmonic Adaptive Latent Component Analysis of Audio and Application to Music Transcription

Benoit Fuentes, *Student Member, IEEE*, Roland Badeau, *Senior Member, IEEE*, and Gaël Richard, *Senior Member, IEEE*

***Abstract*—Recently, new methods for smart decomposition of time-frequency representations of audio have been proposed in order to address the problem of automatic music transcription. However those techniques are not necessarily suitable for notes having variations of both pitch and spectral envelope over time. The HALCA (Harmonic Adaptive Latent Component Analysis) model presented in this article allows considering those two kinds of variations simultaneously. Each note in a constant-Q transform is locally modeled as a weighted sum of fixed narrowband harmonic spectra, spectrally convolved with some impulse that defines the pitch. All parameters are estimated by means of the expectation-maximization (EM) algorithm, in the framework of Probabilistic Latent Component Analysis. Interesting priors over the parameters are also introduced in order to help the EM algorithm converging towards a meaningful solution. We applied this model for automatic music transcription: the onset time, duration and pitch of each note in an audio file are inferred from the estimated parameters. The system has been evaluated on two different databases and obtains very promising results.**

***Index Terms*—Automatic transcription, multipitch estimation, nonnegative matrix factorization, probabilistic latent component analysis.**

I. Introduction

MUSIC transcription consists in describing some physical characteristics of a musical signal by means of some notation system. In many music genres, including western music, a big deal of information is carried on the set of notes played by the instruments, or sources. Classically, a musical note is defined by three attributes: its pitch, duration, and onset time. Automatic music transcription consists in estimating these notes and their attributes given an audio recording (it could also include a clustering of the notes played by a same instrument, but this paper does not address this problem). To this aim, an audio signal is often decomposed into small segments (or frames) on which the pitch of all active notes is estimated. This subtask, called multipitch estimation, is still an open problem, far from being completely solved. For an overview of the proposed methods for this task, the reader is referred to [1]. Recently, a new class of methods has emerged in order to address automatic transcription problems, which consists in modeling a time-frequency representation (TFR) of audio as a sum of basic elements, atoms, or kernels. For instance, an atom can represent the spectrum of a single note, in which case it will be active whenever the corresponding note is played [2]. Such decompositions, called herein TFR factorizations, are widely used in other fields as well, such as source separation [3], main melody estimation or extraction [4], [5] or beat location estimation [6].

TFR factorization techniques can be roughly grouped into three subclasses whether they are unsupervised, supervised or semi-supervised.

Unsupervised factorizations do not use any learning stages but take advantages of redundancies in musical TFR to estimate both the basic atoms and their activations. This is the case of the classic Non-Negative Matrix Factorization (NMF) [2], [7] where each column of a TFR is modeled as a weighted sum of basic spectra (the kernels). The kernels and the time-dependent weights (the activations) are jointly estimated using an iterative algorithm which minimizes some divergence between the input TFR and the factorization model. Many variants and extensions that better take musical signals characteristics into account have been proposed. See for example the shift-invariant model of [8] or the dynamic model with time-frequency activations of [9]. Those techniques are quite appealing since they only rely on the assumption that music is redundant (e.g. no learning stage is needed). It is then theoretically possible to model any kind of musical event, provided it is repeated over time. However, a major drawback with such techniques is that they are underconstrained, and nothing ensures that the decomposition will be sufficiently informative to perform automatic transcription.

An appealing solution to constrain the decomposition is to perform a *supervised factorization*. It consists in using fixed kernels, learned during a training stage, or set manually. For example in [10] or [11], several note kernels are first extracted from monophonic recordings of specific instruments. Then, each column of an input TFR is decomposed using those templates as a dictionary. In [11], the kernels can also be shifted in frequencies so this method is robust to detuning or fundamental frequency modulations. If those methods can be very fast ([10] is a real-time online algorithm), and less sensitive to local minima than unsupervised TFR factorizations, a main drawback is that their performances strongly depend on the

Manuscript received July 19, 2012; revised January 25, 2013 and April 10, 2013; accepted April 15, 2013. Date of publication April 30, 2013; date of current version July 11, 2013. This work was supported in part by the Quaero Programme, funded by OSEO, French State agency for innovation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Laurent Daudet.

The authors are with the Département Traitement du Signal et des Images, Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, 75014 Paris, France (e-mail: benoit.fuentes@telecom-paristech.fr; roland.badeau@telecom-paristech.fr; gael.richard@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2260741

similarity between the fixed kernels and the corresponding spectra of the different sources in the input TFR. Solutions to this problem can be proposed though. For instant, in the model put forward in [5], the spectrum resulting from a linear combination of fixed kernels can be shaped by a smooth filter, so that it better fits the observations.

The approaches of the third subclass are based on a *semi-supervised factorization*. The principle here is to estimate both kernels and activations, and to constrain the kernel to lie in some predefined subspace. This subspace can be defined during a learning stage, as in [12] where the notion of *hierarchical eigeninstruments* is introduced. They represent subspaces, each one of them being used to model a specific class of instruments. Those subspaces are pre-learned and during the factorization stage, one or more eigeninstrument is used to model the note spectra of a source present in the input TFR. The subspace of the kernels can also be set manually. For example, the basic spectra can be modeled as a linear combinations of fixed narrow-band harmonic spectra in order to consider both harmonicity and spectral smoothness of a note spectrum [13]. More simply, the energy of a kernel representing a note can be forced to zero for frequencies between the partials of this note (see [14]). Those methods have many advantages since they ensure both a meaningful factorization and an adaptability to the data.

It is interesting to note that in [12], [13], the hypothesis that the input TFR has redundancies is made. More specifically, it is supposed that a given source in the mixture has note-wise timbre similarities throughout the signal. This supposition reduces the dimension of the pre-defined subspace of the kernels during the factorization step. It is a good assumption for many musical instruments, but not all of them, the human voice being the best counter-example: the timbre depends on the lyrics and not on the sung note. On the opposite, in the model we present in this paper, called Harmonic Adaptive Latent Component Analysis (HALCA), the decomposition is independently performed from one frame to another: no assumption of redundancy is made, and the kernels can stay within all the pre-defined subspace along time. Consequently, it allows considering all kinds of harmonic instruments, having variations of both pitch and spectral envelope over time.

Regardless of the TFR factorization subclass considered, it is also possible to constrain the decomposition by means of the addition of priors or penalty functions. They are applied to the parameters constituting a musical TFR model. Their use can be quite interesting since they act like an inducement for the parameters to converge toward a more likely solution, instead of just constraining the parameters to lie in some subspace. Thus, it is a soft way to add information about the nature of the data to be analyzed. Many priors or penalty terms have been proposed in the literature, such as sparsity or temporal smoothness of the activations [3], [10], [15], [16]. In this paper, three different priors and their derivation are presented: a monomodal prior for monophonic sources (already introduced in [17]), a sparse prior (partially introduced in [18] in a different framework) and a timbre temporal continuity prior.

Before getting to the main contributions and the overview of this paper, it is important to clarify that there are many mathematical frameworks to perform TFR factorization: analytical models [2], [7], [19], [20], probabilistic NMF [21], [22], Probabilistic Latent Component Analysis (PLCA) and its shift invariant version [8], [15], [23], Gaussian Processes [24], Generalized Coupled Tensor Factorization [25], or non-parametric Bayes [26]. In this paper, the framework used is the PLCA, which offers a convenient way to derive convolutive models and introduce priors. In the HALCA model, an input signal is considered as a mixture of several harmonic sources (monophonic or polyphonic) and a noise component. The notes played by the sources can present temporal variations of both spectral envelope and fundamental frequency over time. The system in [5], which models an input spectrogram as the sum of a component of interest and a residual, presents also similar features. Indeed, the component of interest is a source/filter model which can also consider those two kind of variations. In the model of this component, the sources are a set of fixed harmonic kernels, and the filter is a smooth spectral envelope applied to a linear combination of those sources. However, it is more suitable for a single main instrument (especially monophonic) since only one filter is applied to a sum of fixed kernels. Besides, nothing prevents the model of the residual from modeling musical notes as well. Therefore, this system is more adapted to model an input signal as a mixture of a main instrument and an accompaniment. On the opposite, in the HALCA model, each polyphonic source can have its own time varying spectral envelope, and the noise component is designed so that it cannot consider harmonic notes. The main contributions of this paper include the following:

- A generalization of the model presented in [17] for polyphonic signals.
- Besides the two priors of monomodality and sparseness, already introduced respectively in [17] and [18], the introduction of a new prior over the parameters of a PLCA-based model, that enforces temporal continuity.
- For the first time, the HALCA model, as well as the influence of the priors, are evaluated in a task of automatic transcription.

The paper is organized as follows. First, the Constant-Q Transform (CQT), which is the TFR used in this paper, and the PLCA are introduced in Section II. Then, the model we put forward, is explained in Section III. In Section IV, the three priors are presented. Finally, the application to automatic transcription and the evaluation are described in Section V, before concluding in Section VI.

II. Tools and Framework

A. Constant-Q Transform

The system put forward in this paper performs the analysis in the time-frequency domain. Thus, a constant-Q transform (CQT) [27] is first applied to the audio signal to be analyzed. The CQT is a complex time-frequency representation of a temporal signal, with a logarithmic frequency scale. This characteristic offers a major advantage for musical signals: the spacing between two given partials of a harmonic note remains the same, regardless of its pitch. A change of fundamental frequency can thus be seen as a frequency shift of the partials, and it is possible to devise shift-invariant models, such as the HALCA model pre-

sented in this paper. This model takes as input non-negative data, therefore, we apply a positive transformation to the CQT of a signal. Actually, every TFR in this paper are calculated the same way. First, the complex CQT X_{ft} (f and t being frequency and time indexes) of a monophonic signal, with 3 bins/semitones and a time step of 10 ms, is calculated for f from 27.5 Hz to 7040 Hz (we used the implementation provided in [28]). Then the input data V_{ft} is computed as follows: $V_{ft} = \sqrt{|X_{ft}|}$. Using the square root is equivalent to applying a slight compression on the coefficients, and experiments have shown that the analysis model we present here gives better results that way. By abuse of language, the term *CQT* will refer to this kind of TFR from now on.

B. Introduction to Probabilistic Latent Component Analysis

PLCA [23] is a tool for non-negative data analysis (in this paper, the data are the non-negative coefficients that compose the magnitude V_{ft} of the CQT of a signal). The observations V_{ft} are modeled as the histogram of the sampling of J independent and identically distributed random variables (f_j, t_j) (j from 1 to J is the number of the draw) that correspond to time-frequency bins. They are distributed according to the probability distribution $P(f,t;\Lambda)$, Λ being a set of parameters. The way $P(f,t;\Lambda)$ is structured induces the desired decomposition of V_{ft}: Λ is estimated by maximizing the log-likelihood of the observations given the parameters, or, if there is a prior distribution $Pr(\Lambda)$ of the parameters, the posterior log-probability. One can calculate the log-likelihood function (we denote $\bar{x}$ the set of variables $\{x_j\}_{j=1\ldots J}$):

$$\begin{aligned} L_\Lambda(\bar{f},\bar{t}) &= \ln\left(P(\bar{f},\bar{t};\Lambda)\right) \\ &= \ln\left(\prod_j P(f_j,t_j;\Lambda)\right) \\ &= \sum_j \ln\left(P(f_j,t_j;\Lambda)\right) \\ &= \sum_j \sum_{f,t} \mathbb{1}_{(f_j,t_j)}(f,t)\ln\left(P(f,t;\Lambda)\right) \end{aligned}$$

where $\mathbb{1}_y(x)$ is the indicator function. That leads to:

$$L_\Lambda(\bar{f},\bar{t}) = \sum_{f,t} V_{ft}\ln\left(P(f,t;\Lambda)\right), \tag{1}$$

since V_{ft} is considered as a histogram. The posterior log-probability is then given by (up to an additive constant with respect to Λ):

$$\ln\left(P(\Lambda|\bar{f},\bar{t})\right) = L_\Lambda(\bar{f},\bar{t}) + \ln\left(Pr(\Lambda)\right). \tag{2}$$

In the basic PLCA model, a latent variable z is introduced, f and t are conditionally independent given z, and $P(f,t;\Lambda)$ is modeled as:

$$P(f,t;\Lambda) = \sum_z P(z)P(f|z)P(t|z) = \sum_z P(z,t)P(f|z),$$

where Λ is the set of parameters $\{P(z,t), P(f|z)\}_{z,t,f}$. $P(f|z)$ then corresponds to several basic spectra and $P(z,t)$ to their time activations, similarly to the classic NMF.

In the shift-invariant version of PLCA [8], as well as in the HALCA model, f results from the sum of two latent random variables, and $P(f,t;\Lambda)$ is consequently the convolution of two probability distributions, as we will see in next section. Since there is usually no closed-form solution for the maximization of the log-likelihood or the posterior, the Expectation-Maximization (EM) algorithm is used to estimate the model parameters. In this document, $\mathbb{Z}$ and $\mathbb{R}$ will refer to the sets of integers and real numbers, $]a,b[$ to the set $\{x \in \mathbb{R} | a < x < b\}$ and $[[a,b]]$ to the set $\{x \in \mathbb{Z} | a \leq x \leq b\}$.

III. HALCA Model

A. Model Representation

Let us introduce a first latent variable c in order to decompose the CQT of a musical signal as the sum of a polyphonic harmonic signal (in this case, $c = h$) and a noise signal ($c = n$) (the notations $P_h(.)$ and $P_n(.)$ are used for $P(.|c = h)$ and $P(.|c = n)$):

$$P(f,t) = P(c=h)P_h(f,t) + P(c=n)P_n(f,t), \tag{3}$$

where $P_h(f,t)_{(f,t)\in\mathbb{Z}\times[[1,T]]}$ and $P_n(f,t)_{(f,t)\in\mathbb{Z}\times[[1,T]]}$ respectively represent the CQTs of the polyphonic and the noise signal. $P(c)_{c=h,n}$ corresponds to the normalized global energy of each part. We can now consider the polyphonic part as the sum of S sources (designated by the latent variable s), each of them supposed to represent a single instrument:

$$P_h(f,t) = \sum_s P_h(f,t,s). \tag{4}$$

Each column of $P_h(f,t,s)$ then represents the spectrum of one or more harmonic notes, played by a single instrument (we suppose that all instruments are harmonic). Let us see now how each column of $P_h(f,t,s)$ and $P_n(f,t)$ is modeled.

1) Instrument Model: In order to account for the non-stationary nature of many musical instruments in terms of both pitch and spectral envelope, the model used for one instrument is the same as in [17]. This model allows considering simultaneously those two kinds of non-stationarities. At time t, the spectrum of source s, represented by $P_h(f,t,s)$, is decomposed as a weighted sum of Z fixed narrow-band harmonic spectral kernels, or kernel distributions, denoted $P_h(\mu|z)_{(\mu,z)\in[[1,F]]\times[[1,Z]]}$, spectrally convolved by a time-frequency impulse distribution $P_h(i,t,s)_{(i,t)\in[[1,I]]\times[[1,T]]}$ (f is then defined as the sum of the two random variables μ and i):

$$P_h(f,t,s) = \sum_{z,i} P_h(i,t,s)P_h(f-i|z)P_h(z|t,s). \tag{5}$$

The parameters have the following characteristics:

- all kernels $P_h(\mu|z)$ share the same fundamental frequency, but have their energy concentrated at a given harmonic (their design will be discussed later in Section III-A-4),
- the weights applied to the kernel distributions, denoted $P_h(z|t,s)$ and called envelope coefficients, define the spectral envelope of the notes played at time t by source s,

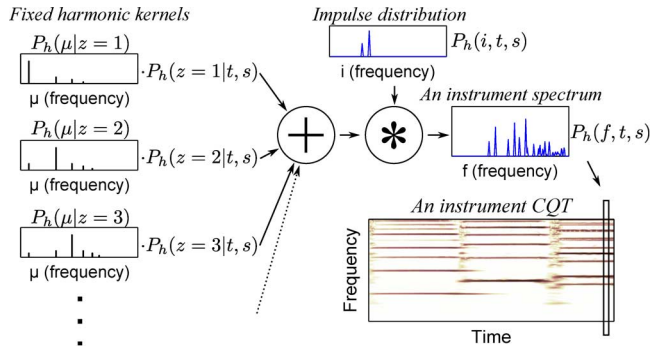


Fig. 1. Spectrum model for source s at time t. Each kernel has its main energy concentrated on a given harmonic (multiple of a reference fundamental frequency), and the rest of the energy is shared between adjacent partials.

- each column of the impulse distribution $P_h(i,t,s)_i$ can be unimodal or multimodal, each mode corresponding to the pitch of one note,
- in order to ensure that the model can fit any spectral spreading of the partials (for instance, a continuous variation of pitch induces a larger spreading at a given time), kernels have energy only for the frequency bins corresponding at harmonics and the spreading is taken into account in the impulse distribution.

Fig. 1 illustrates this instrument model. In this model, the spectral shape of each note can evolve over time. However, we need to suppose that at given t and s, all simultaneous notes of a single source have the same spectral shape. This hypothesis is not very realistic for real music instruments, but it does not appear to be critical for the application to music transcription, as we will observe in Section V. The following remark can help to understand why. In practice, one source does not necessarily represent a specific instrument: several sources can be used to model a single real instrument, and one source can contribute to the modeling of several notes played by several instruments. Thus, we could say that a source represents a "meta-instrument." For automatic transcription, as we mean it (i.e. with no clustering of the notes with respect to the instruments), this characteristic is not a problem since only the global impulse distribution, defined as $P_h(i,t) = \sum_s P_h(i,t,s)$ will be used to estimate the onset, offset and pitch of each note.

2) Noise Model: In order to consider smooth structures in a CQT, the noise is modeled as the convolution of a fixed smooth narrowband window, $P_n(\mu)$ and a noise impulse distribution $P_n(i,t)$ (the term impulse is no longer relevant in this case, but we keep it for the sake of consistency). The model can be written:

$$P_n(f,t) = \sum_i P_n(i,t) P_n(f-i) \tag{6}$$

It is illustrated in Fig. 2.

3) Complete Model: By grouping (3), (4), (5) and (6), we can formulate the complete HALCA model:

$$P(f,t;\Lambda) = P(c=h) \sum_{s,i,z} P_h(i,t,s) P_h(f-i|z) P_h(z|t,s) + P(c=n) \sum_i P_n(i,t) P_n(f-i), \tag{7}$$

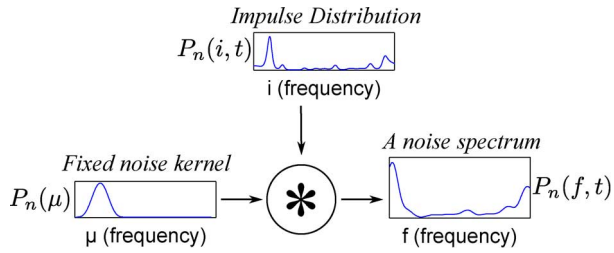


Fig. 2. Noise spectrum model.

TABLE I
PARAMETERS OF THE HALCA MODEL

Parameters	Semantic definition
$P(c=h)$	Relative energy of the polyphonic harmonic component.
$P_h(i,t,s)$	Time-frequency activations of each source.
$P_h(\mu\|z)$	z^{th} fixed narrowband kernels.
$P_h(z\|t,s)$	Weights of the kernels at time t for source s (defines the spectral enveloppe).
$P(c=n)$	Relative energy of the noise component.
$P_n(i,t)$	Time-frequency distribution of noise.
$P_n(\mu)$	Smooth narrowband noise kernel (fixed).

where $\Lambda = \{P(c), P_h(i,t,s), P_h(z|t,s), P_n(i,t)\}_{i,t,s,z,c}$ is the set of parameters to be estimated. In Table I, the parameters of the HALCA model are listed, as well as their semantic meanings.

From Λ, useful pieces of information can then be deduced. For instance, the frame by frame pitch activity can be extracted from the overall impulse distribution $P_h(i,t) = \sum_s P_h(i,t,s)$, so we can automatically perform the transcription (see Section V).

4) Designing the Kernels: In the HALCA model, we aim at modeling the spectrum of any real harmonic note as a linear combination of several basis vectors called kernels. Due to the convolutive nature of the model, the kernels can be designed independently of the pitch of a note and the spectral spreading of its partials. Many approaches are possible to design such kernels. One possibility is to define single harmonic kernels (e.g. each kernel has energy only on a given harmonic) similarly to the HTC model [20], where a harmonic spectrum is modeled as the sum of Gaussians representing the partials. This allows considering any kind of spectral envelope. However, if no prior on the parameters is added, it can easily lead to octave errors since a note with a f_0 fundamental frequency could be modeled by a note with a $f_0/2$ fundamental frequency in which all odd harmonics are zero. To avoid octave errors, it is possible to use a fewer number of kernels in which several adjacent partials have non-zero energy, like in [13]. Nevertheless, using such kernels is not appropriate for modeling notes with missing partials.

A more satisfying choice is obtained as a trade-off between those two possibilities: the number of kernels is set to the maximum number of partials we consider and each kernel has its main energy concentrated on a given harmonic, the rest of the

energy being shared between few adjacent partials. The kernels are defined as follows: $\forall z \in [[1,16]]$,

$$P(\mu|z) = \begin{cases} W(0) + 10 + \sum_{j=-3}^{-z} W(j) \text{ if } \mu = \mu_z, \\ W(j) \text{ if } \mu = \mu_{z+j} \text{ with } j \in [[-3,3]] \setminus \{0\}, \\ \qquad j \leq 16 - z \text{ and } j \geq 1 - z \\ 0 \text{ else,} \end{cases}$$

where $W(j)_{j\in[[-3,3]]}$ is a symmetric Hamming window centered at 0 and where $\mu_1, \mu_2, \cdots \mu_{16}$ are the rounded theoretical frequencies of the 16 first partials of a harmonic spectrum of fundamental frequency $\mu_1 = 1$.

B. EM Algorithm and Update Rules

Given an observed CQT V_{ft} (we suppose that $V_{ft} = 0$ for $f \notin [[1,F]]$) and a fixed set of kernels $P_h(\mu|z)$, the purpose is to find the set of distributions $\Lambda = \{P(c), P_h(i,t,s), P_h(z|t,s), P_n(i,t)\}$ which maximizes the likelihood of the observations given the parameters (for now, we suppose there is no prior). The EM algorithm defines update rules for the parameters so that the likelihood of the observations $L_\Lambda(\bar{f},\bar{t})$, given by (1), is not decreasing at any iteration.

In the HALCA model, variables f and t are the observations whereas i, s, c and z are latent variables. It can be shown that the conditional expectation of the joint log-likelihood $\ln(P(\bar{f},\bar{t},\bar{i},\bar{s},\bar{c},\bar{z}))$ given the observations and the parameters is given by:

$$\begin{aligned} Q_\Lambda = & \sum_{i,s,z} V_{ft} P(i,s,z,c=h|f,t) \\ & \times [\ln(P(c=h)) + \ln(P_h(z|t,s)) \\ & \quad + \ln(P_h(f-i|z)) + \ln(P_h(i,t,s))] \\ & + \sum_i V_{ft} P(i,c=n|f,t) \\ & \times [\ln(P(c=n)) + \ln(P_n(i,t)) \\ & \quad + \ln(P_n(f-i))]. \end{aligned} \tag{8}$$

In the expectation step, the a posteriori distributions of the latent variables are computed using Bayes' theorem:

$$\begin{aligned} & P(i,s,z,c=h|f,t) \\ & \quad = \frac{P(c=h)P_h(i,t,s)P_h(f-i|z)P_h(z|t,s)}{P(f,t;\Lambda)}, \end{aligned} \tag{9}$$

$$\begin{aligned} & P(i,c=n|f,t) \\ & \quad = \frac{P(c=n)P_n(i,t)P_n(f-i)}{P(f,t;\Lambda)}, \end{aligned} \tag{10}$$

$P(f,t;\Lambda)$ being defined by (7).

Then, in the expectation step, Q_Λ is maximized with respect to (w.r.t.) Λ under the constraint that all probability distributions sum to one. This leads to the following update rules:

$$P(c=h) \propto \sum_{f,t,i,s,z} V_{ft} P(i,s,z,c=h|f,t), \tag{11}$$

$$P_h(i,t,s) \propto \sum_{f,z} V_{ft} P(i,s,z,c=h|f,t), \tag{12}$$

$$P_h(z|t,s) \propto \sum_{f,i} V_{ft} P(i,s,z,c=h|f,t), \tag{13}$$

$$P(c=n) \propto \sum_{f,t,i} V_{ft} P(i,c=n|f,t), \tag{14}$$

$$P_n(i,t) \propto \sum_{f} V_{ft} P(i,c=n|f,t). \tag{15}$$

The EM algorithm first consists of initializing Λ, then iterating (9) and (10), the various update rules ((11), (12), (13), (14), (15)) and finally the normalization of every parameter so that the probabilities sum to one.

IV. Use of Priors

The HALCA model can fit any harmonic instrument with time-varying pitch and spectral envelope but it has some drawbacks. First, it is not identifiable since different sets of parameters can explain a given observation. For example, if the input is a single harmonic note, nothing in the HALCA model prevents from modeling it as a sum of several notes whose pitches correspond to the different harmonics. Moreover, the model does not take into account some temporal continuities of acoustical characteristics of musical signals: not only useful information is missed, but also the algorithm could converge to an irrelevant solution.

In order to get around those issues, an option is to constrain the model parameters to stay within some subset, as proposed in [20], where the power envelope of each partial in a note is parameterized as a sum of Gaussian windows regularly spaced over time. Similarly, in order to force the columns of the impulse distribution to be monomodal in the case where each source is a monophonic instrument, an attracting approach would consist in parameterizing them as Gaussian distributions. However, we noticed a major problem with such solutions: since we prevent the parameters from moving away from a given subset during the iterations, the algorithm becomes much more sensitive to local minima.

A second option is to add informative priors over the parameters. This solution is commonly used in the literature (see [8], [16], [20]) and has the advantage of being more flexible. In this section, we propose three different priors and evaluate their merit.

A. Monomodal Prior

We consider here the case where each source is monophonic: at each time frame, source s plays a single note. Ideally, after convergence of the EM algorithm, for given t and s, the estimated vector $P_h(i,t,s)$ would be a monomodal vector and the value of the mode would give the pitch of source s at time t. Unfortunately, this is not necessarily the case, and in practice we could end up with a multimodal vector, the modes corresponding to the different partials of the played note. Since we would like to keep only the mode of lowest frequency, we employ a monomodal prior that forces those vectors to have both low variance and low mean. To do so, the HALCA model needs to be adapted: the impulse distribution is decomposed as

$$P_h(i,t,s) = P_h(t,s)P_h(i|t,s) \tag{16}$$

where $P_h(t,s)$ and $P_h(i|t,s)$ respectively represent the energy of instrument s at time t and the corresponding pitch distribution. If no prior was added, (12) would become the set of equations

$$P_h(i|t,s) \propto \sum_{f,z} V_{ft} P(i,s,z,c=h|f,t) \tag{17}$$

$$P_h(t,s) \propto \sum_{f,z,i} V_{ft} P(i,s,z,c=h|f,t) \tag{18}$$

and then the normalization would be performed. The monomodal prior we put forward is applied on each distribution $P_h(i|t,s)$ at every time frame and for every source. It is based on an adequate measure that we call asymmetric variance, introduced for the first time in [17] (for the sake of simplicity, we fix a given source s and a given time t, and we define $\boldsymbol{\theta}$ as the vector of coefficients $\theta_i = P_h(i|t,s)$):

$$\begin{aligned}\mathrm{avar}_\gamma(\boldsymbol{\theta}) &= \sum_i \left(e^{\gamma i} - e^{\gamma \sum_{\tilde{i}} \tilde{i}\theta_{\tilde{i}}}\right)\theta_i \\ &= \left(\sum_i e^{\gamma i}\theta_i\right) - e^{\gamma \sum_i i\theta_i} \text{ since } \sum_i \theta_i = 1. \end{aligned} \tag{19}$$

This measure depends on the hyperparameter $\gamma > 0$ which defines the strength of the asymmetry. It can be proven, due to the strict convexity of the exponential function, that

$$\mathrm{avar}_\gamma(\boldsymbol{\theta}) \geq 0,$$

and

$$\mathrm{avar}_\gamma(\boldsymbol{\theta}) = 0 \Leftrightarrow \exists i_0 | \forall i, \theta_i = 1 \text{ if } i = i_0 \text{ and } 0 \text{ otherwise.}$$

In order to force $\mathrm{avar}_\gamma(\boldsymbol{\theta})$ to have a low value during the training, the following prior distribution is introduced:

$$Pr(\boldsymbol{\theta}) \propto \exp\left(-\alpha \mathrm{avar}_\gamma(\boldsymbol{\theta})\right) \tag{20}$$

where $\alpha > 0$ is a hyperparameter indicating the strength of the prior. The maximization step is now replaced by a maximum a posteriori (MAP) step, meaning that instead of maximizing Q_Λ, we maximize $Q_\Lambda + \ln(Pr(\boldsymbol{\theta}))$ w.r.t. the model parameters. Only the update rule for $P_h(i|t,s)$ changes. Maximizing the posterior probability w.r.t. $P_h(i|t,s)$ amounts to maximizing on $\Omega =]0,1[^I$ the following functional under the constraint $\sum_i \theta_i = 1$:

$$\begin{aligned}\mathcal{M} : \Omega =]0,1[^I &\longrightarrow \mathbb{R} \\ \boldsymbol{\theta} &\longmapsto \sum_i w_i \ln(\theta_i) - \alpha\left(\sum_i e^{\gamma i}\theta_i\right) \\ &\quad + \alpha e^{\gamma \sum_i i\theta_i}, \end{aligned} \tag{21}$$

where $w_i = \sum_{f,z} V_{ft} P(i,s,z,c=h|f,t)$. If $\hat{\boldsymbol{\theta}}$ is the maximum that we are looking for, then according to the Karush-Kuhn-Tucker (KKT) conditions [29], there exists an unique $\upsilon \in \mathbb{R}$ such that

$$\forall i \in \mathbb{Z}, \hat{\theta}_i = \frac{w_i}{\alpha\left(e^{\gamma i} - \gamma i e^{\gamma \sum_{\tilde{i}} \tilde{i}\hat{\theta}_{\tilde{i}}}\right) + \upsilon}. \tag{22}$$

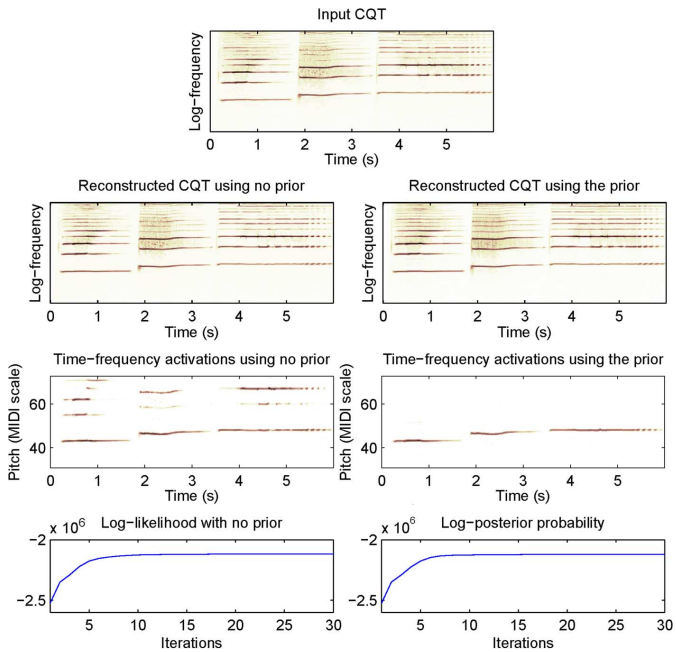


Fig. 3. Illustration with a single source ($S = 1$) of the use of the monomodal prior: if the estimated CQT remains almost unchanged, time-frequency activations, defined here as $P_h(t, s = 1)P_h(i|t, s = 1)$, become monomodal for each time frame. In both cases, the convergence criterion increases over the iterations. The input signal corresponds to three successive notes played by a harmonica.

Unfortunately, there is no closed-form solution for θ_i, and we cannot be sure that there is a unique solution. However, numerical simulations showed that the fixed point Algorithm 1 always converges to a solution that makes the posterior probability increase from one iteration of the EM algorithm to the next. This algorithm is independently run for every value of t and s. Fig. 3 illustrates the effect of the prior, along with the growth of the posterior probability over the iterations. When the number S of monophonic sources is known and low ($S = 1, 2$), the monomodal prior appears to be effective, as proven in Section V-B. However, when it comes to higher levels of polyphony, the use of this prior leads to irrelevant solutions, due to a too large number of local minima. Besides, the case where all sources are monophonic is quite restrictive, and one would like to deal with polyphonic instruments such as the guitar. In next section, an alternative prior that encourages sparsity on the impulse distribution is introduced.

Algorithm 1: Fixed-point method for the monomodal prior

$\forall i \in [[1, I]], \hat{\theta}_i \leftarrow w_i / \sum_{\tilde{i}} w_{\tilde{i}}$;

repeat

$\cdot m \leftarrow \sum_i i\hat{\theta}_i$;

$\cdot \forall i \in [[1, I]], c_i \leftarrow \alpha(e^{\gamma i} - \gamma i e^{\alpha m})$;

$\cdot$ find the unique υ such that $\sum_i w_i/(c_i + \upsilon) = 1$ and $\forall i, w_i/(c_i + \upsilon) \geq 0$ (we used Laguerre's method [30]);

$\cdot \forall i \in [[1, I]], \hat{\theta}_i \leftarrow w_i/(c_i + \upsilon)$;

until *convergence*

B. Sparse Prior

In order to account for polyphonic instruments, the monomodal prior we presented in previous section can be replaced with a sparsity constraint on the impulse distribution $P_h(i,t,s)$. If we consider P_{Ih} as a single long vector $\boldsymbol{\theta}$ of coefficients θ_j (with $j \in [[1,J]]$ where $J = I \times T \times S$), enforcing its global sparsity (which is, in a way, equivalent to saying that a musical score is a sparse representation) allows considering several levels of sparsity:

- few notes are present at a given time frame,
- few sources contribute to the production of a given note,
- a given source is not necessarily active at every moment.

Several solutions have been suggested in the literature in order to enforce sparsity in the framework of PLCA. In [15] for instance, an exponentiated negative-entropy term is used as a prior on parameter distributions. However, the solution to the maximization step involves complex transcendental equations and resolving them sometimes leads to numerical errors. Moreover, the resulting posterior is not a concave function, and the corresponding Lagrange function may have more than one stationary point: we notice in practice that the proposed fixed point algorithm in [15] does not necessarily converge towards the global maximum during the M-step. An alternative is presented in [31] where a power greater than 1 is applied to the distribution that one wants to make sparser, just before the normalization in the M-step. It is indeed an easy way to enforce sparsity, but it does not rely on theoretical results, and nothing proves that the likelihood is still increasing with such update rules.

We put forward the following sparsity prior on $\boldsymbol{\theta}$, firstly introduced in [18]:

$$Pr(\boldsymbol{\theta}) \propto \exp\left(-2\beta\sqrt{J}\,\|\boldsymbol{\theta}\|_{\frac{1}{2}}\right), \tag{23}$$

where $\|\boldsymbol{\theta}\|_{1/2} = \sum_j \sqrt{\theta_j}$. β is a positive hyperparameter indicating the strength of the prior and the constant $\sqrt{J}$ is used so that the strength is independent of the size of the data. The new update rule for θ_j is obtained by maximizing $Q_\Lambda + \log(Pr(\boldsymbol{\theta}))$ w.r.t. $\boldsymbol{\theta}$, which amounts to maximizing on $\Omega =]0,1[^J$ the following functional under the constraint $\sum_j \theta_j = 1$:

$$\begin{aligned} \mathcal{S} : \Omega =]0,1[^J &\longrightarrow \mathbb{R} \\ \boldsymbol{\theta} &\longmapsto \sum_j w_j \log(\theta_j) - 2\beta\sqrt{J}\sum_j \sqrt{\theta_j}, \end{aligned} \tag{24}$$

where $\{w_j\}_{j\in[[1,J]]} = \{\sum_{f,z} V_{ft} P(i,s,z,c = h|f,t)\}_{i,t,c}$. In Appendix A, it is proven that if $\beta^2 < \sum_j w_j^2/J$, which is always the case in practice, then the argument of this maximum is given by:

$$\forall j, \theta_j = \frac{2w_j^2}{J\beta^2 + 2\rho w_j + \beta\sqrt{J}\sqrt{J\beta^2 + 4\rho w_j}}, \tag{25}$$

where ρ is the unique positive real number such that $\sum_j \theta_j = 1$. It can be found with any numerical root finding algorithm (we used the *fzero.m* Matlab function). In Fig. 4, the effect of the sparse prior and the growth of the criterion over the iterations can be observed.

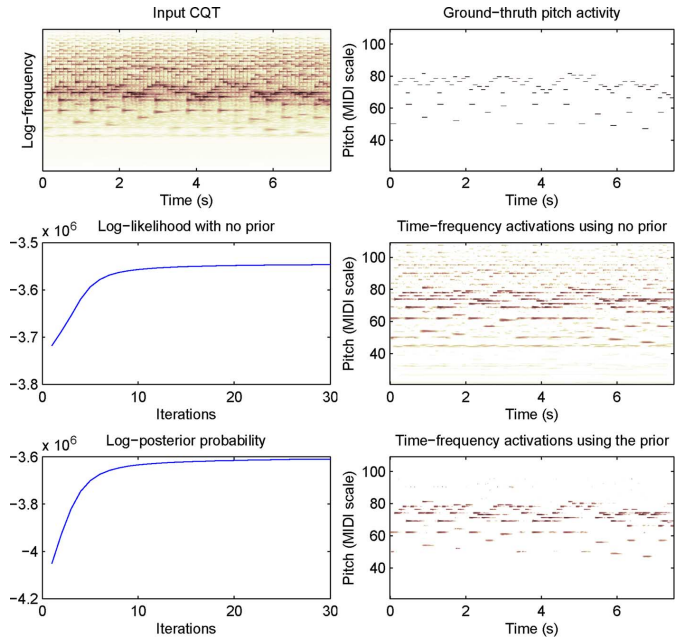


Fig. 4. Illustration of the use of the sparse prior and the growth of the criterion over the iterations. The input signal is an 8 s excerpt from Bach's *Prelude and Fugue in D major BWV 850* and the HALCA model has been estimated with $S = 2$ sources. The time-frequency activations correspond to the summation over the sources of the impulse distribution: $\sum_s P_h(i,t,s)$.

C. Spectral Envelope Temporal Continuity Prior

The HALCA model allows modeling notes with time-varying fundamental frequencies or spectral envelopes. But so far, nothing constrains those attributes to evolve smoothly over time, whereas it could be a useful piece of information, as usually observed in real-world musical sounds. Many solutions have been suggested in order to enforce temporal continuity whether in the framework of NMF or PLCA. One can mention for instance [3] or [16], where a smoothness constraint is imposed respectively via a penalty term in the NMF cost function and via a prior on the parameters in the framework of Bayesian NMF. In [8], the temporal continuity is imposed by applying a Kalman filter smoothing on the impulse distributions between two iterations of the EM algorithm. The common characteristic of most of the proposed methods for enforcing the continuity of spectrogram decomposition is that they enforce the temporal smoothness of the energy of the sources. The new idea here is instead to enforce the temporal continuity of the timbre of the sources, which in the HALCA model is represented by the envelope coefficients $P_h(z|t,s)$.

We use this constraint for several reasons. First, it can prevent the EM algorithm from staying stuck in a local maximum. Then, contrary to a prior of energy temporal smoothness, enforcing timbre temporal continuity does not disfavor hard attacks of notes. A more obvious reason is that we already introduced the sparse prior and the monomodal prior applied on the impulse distribution (Sections IV-A and IV-B), and two different priors on the same set of parameters might lead to some difficulties, in terms of mathematical calculation.

For a given source s, let $\boldsymbol{\Theta}$ and $\boldsymbol{W}$ respectively denote the $Z \times T$ matrix of coefficients $\theta_z^t = P_h(z|t,s)$ and the $Z \times T$

matrix of coefficients $w_z^t = \sum_{f,i} V_{ft} P(i,s,z,c=h|f,t)$. We introduce a new prior on $\mathbf{\Theta}$, defined as:

$$Pr(\mathbf{\Theta}) \propto \left(\prod_z \prod_{t=2}^{T} 2 \frac{\sqrt{\theta_z^t \theta_z^{t-1}}}{\theta_z^t + \theta_z^{t-1}} \right)^{\chi} \tag{26}$$

where χ is a positive hyperparameter indicating the strength of the prior. Such a prior indeed favors slow evolution of the coefficients in each row of $\mathbf{\Theta}$, since the closer two numbers are, the bigger the ratio between their geometric and arithmetic means is. Maximizing $Q_\Lambda + \ln(Pr(\mathbf{\Theta}))$ w.r.t. $\mathbf{\Theta}$ corresponds to maximizing the function:

$$\begin{aligned} \mathcal{T} : \Omega =]0,1[^{Z \times T} &\longrightarrow \mathbb{R} \\ \mathbf{\Theta} &\longmapsto \sum_z \sum_{t=1}^{T} w_z^t \log\left(\theta_z^t\right) \\ &\quad + \chi \sum_z \sum_{t=2}^{T} \ln\left(\frac{\sqrt{\theta_z^t \theta_z^{t-1}}}{\theta_z^t + \theta_z^{t-1}} \right), \end{aligned} \tag{27}$$

under the constraint $\forall t, \sum_z \theta_z^t = 1$. If $\hat{\mathbf{\Theta}}$ is the maximum that we are looking for, then according to the KKT conditions, $\forall t$ there exists a unique $\sigma_t \in \mathbb{R}$ such that

$$\forall z \begin{cases} \hat{\theta}_z^t = \frac{w_z^t + \chi}{\sigma_t + \frac{\chi}{2\hat{\theta}_z^t} + \frac{\chi}{\hat{\theta}_z^{t+1} + \hat{\theta}_z^t}} & \text{if } t = 1 \\ \hat{\theta}_z^t = \frac{w_z^t + \chi}{\sigma_t + \frac{\chi}{\hat{\theta}_z^{t-1} + \hat{\theta}_z^t} + \frac{\chi}{\hat{\theta}_z^{t+1} + \hat{\theta}_z^t}} & \text{if } t \in [[2, T-1]] \\ \hat{\theta}_z^t = \frac{w_z^t + \chi}{\sigma_t + \frac{\chi}{\hat{\theta}_z^{t-1} + \hat{\theta}_z^t} + \frac{\chi}{2\hat{\theta}_z^t}} & \text{if } t = T \end{cases} . \tag{28}$$

Unfortunately, as for the monomodal prior, there is no closed form solution, and we cannot be sure that there is a unique solution. However, numerical simulations showed that the fixed point Algorithm 2 always converges to a solution that makes the posterior probability increase from one iteration of the EM algorithm to the next. Fig. 5 illustrates the use of the prior.

Algorithm 2: Fixed-point method for the temporal continuity prior

$\forall (z,t) \in [[1,Z]] \times [[2,T]], \hat{\theta}_z^t \leftarrow w_z^t / \sum_{\tilde{z}} w_{\tilde{z}}^t$;

repeat

$\cdot \forall z \in [[1,Z]], s_z^1 \leftarrow \chi / (2\hat{\theta}_z^1)$;

$\cdot \forall (z,t) \in [[1,Z]] \times [[2,T]], s_z^t \leftarrow \chi / (\hat{\theta}_z^{t-1} + \hat{\theta}_z^t)$;

$\cdot \forall z \in [[1,Z]], s_z^{T+1} \leftarrow \chi / (2\hat{\theta}_z^T)$;

$\cdot \forall t \in [[1,T]]$, find the unique σ^t such that $\sum_z (w_z^t + \chi)/(\sigma^t + s_z^t + s_z^{t+1}) = 1$ and $\forall z, (w_z^t + \chi)/(\sigma^t + s_z^t + s_z^{t+1}) \geq 0$ (we used Laguerre's method [30]);

$\cdot \forall (z,t) \in [[1,Z]] \times [[2,T]], \hat{\theta}_z^t \leftarrow (w_z^t + \chi)/(\sigma^t + s_z^t + s_z^{t+1})$;

until *convergence*;

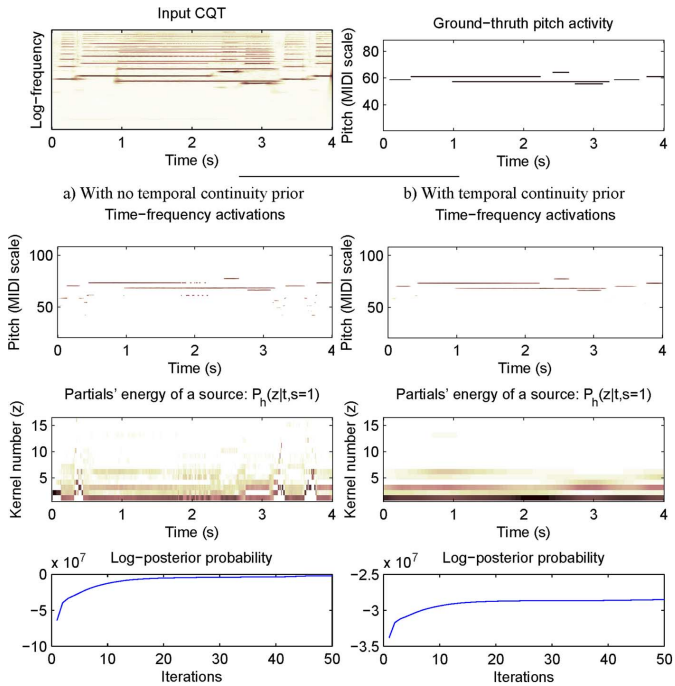


Fig. 5. Illustration of the use of the temporal continuity prior and the growth of the criterion over the iterations. The input signal corresponds to several notes of clarinet and horn. The HALCA model has been estimated with $S = 2$ sources. The time-frequency activations correspond to the summation over the sources of the impulse distribution: $\sum_s P_h(i,t,s)$. The monomodal prior is used in both cases.

V. Application and Evaluation

A. Some Useful Remarks

1) Parameters Initialization: As for many iterative maximization algorithms, the way parameters are initialized has a very important effect on the convergence to a local maximum. After experimentation, the following recommendations can be provided:

- random initialization is not recommended,
- a good initialization of the impulse distribution is the uniform distribution $(P_h(i,t,s) = 1/(I \times T \times S))$,
- envelope coefficients $P_h(z|t,s)$ must be initialized differently for the different sources s.

2) Defining the Hyperparameters: Each prior mentioned in Section IV depends on one or more hyperparameter (for instance the hyperparameter β in the case of the sparse prior). For now, no research has been made on how to automatically estimate their value, hence the need to manually predefine them. However, after experimentation, we noticed that a good way to avoid local minima was to increase the values of the hyperparameters from 0 to the predefined values during the first iterations of the EM algorithm. This strategy stands for the monomodal and the sparse priors. Concerning the temporal prior, the value of χ can be fixed from the first iteration.

B. Monopitch Estimation

To evaluate the relevance of the HALCA model, it has first been tested on a task of monopitch estimation. The database used for the evaluation consists of 3307 isolated notes from the Iowa database [32]. It includes recordings of several instruments, playing over their full range, and with various play

TABLE II
DB_{train}: LIST OF THE AUDIO EXCERPTS FROM RWC CLASSICAL GENRE DATABASE [35] AND THEIR CORRESPONDING NUMBER (#) OF SOURCES AND POLYPHONY (POL.) LEVEL

Symbol	Title (Composer)	Catalog number	# of instruments	Pol. level (Ave. / Max.)
rwc(1)	The Musical Offering (Bach)	RWC-MDB-C-2001 No. 12	2	1.3 / 4
rwc(2)	String Quartet No. 19 (Mozart)	RWC-MDB-C-2001 No. 13	4	2.7 / 5
rwc(3)	Clarinet Quintet Op.115. (Brahms)	RWC-MDB-C-2001 No. 17	5	3.8 / 10
rwc(4)	Horn Trio Op.8 (Brahms)	RWC-MDB-C-2001 No. 18	3	4.2 / 10
rwc(5)	The Anna Magdalena Bach Notebook (Bach)	RWC-MDB-C-2001 No. 24a	1	1.9 / 5

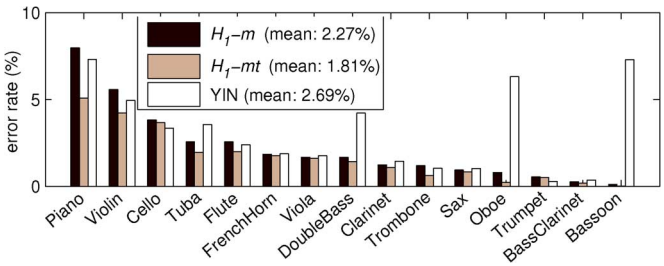


Fig. 6. Simulation results: averaged error rates for each instrument of the database in a task of monopitch estimation. The error rate corresponds to the proportion of frames for which the MIDI pitch is wrongly estimated. The monomodal prior is used for both HALCA algorithms.

modes and nuances. The CQT of each audio file is computed as described in Section II-A. It is then analyzed with two different versions of the HALCA algorithms, with a number of sources set to $S = 1$: one using the monomodal prior only (denoted $H_1 - m$), one using both the monomodal and the temporal continuity priors (denoted $H_1 - mt$). The pitch is finally inferred for each time frame from the maximum of the impulse distribution $P_h(i|t, s = 1)$ and associated to the closest MIDI note. The methods are compared with the YIN algorithm [33], available on the author's website[1], using 100ms time frames. The output of this algorithm is as well rounded to the closest MIDI pitch. Results are illustrated in Fig. 6. Several conclusions can be made from those results. First, we can see that our results are comparable to those of YIN algorithm, except for the oboe and the bassoon. Those two instruments have indeed very high spectral centroids, regardless of the pitch, and where YIN makes upper octave or twelfth errors, the HALCA model adapts itself to the spectral shapes. It is also shown that the addition of the temporal prior to the monomodal prior improves the performances for each instrument. Since the HALCA model seems to be relevant for any kind of instrument, it can now be applied to polyphonic music.

C. Automatic Transcription

This section is dedicated to the application of the HALCA model to automatic music transcription. First, the complete transcription system is presented as well as the metrics used to assess its performance. Then, the model is tested on a first database, on which several versions of our algorithm are compared, including different sets of values for the hyperparameters and different values for the model order S. Finally, three versions of the HALCA model are compared to other state of the art transcription algorithms on a second database.

1) Description of the Transcription System:

- For a given input audio signal, the CQT is first calculated as described in Section II-A.
- It is then analyzed with the HALCA algorithm and the time-frequency activation matrix $P_h(i,t)$ is deduced from the estimated impulse distribution: $P_h(i,t) = \sum_s P_h(i,t,s)$.
- To obtain a pitch activity matrix A, $A(p,t)$ corresponding to the velocity of pitch p (integer on the MIDI scale) at time t, every peak of each vector $(P_h(i,t))_i$ (t from 1 to T) is first detected. Then, at time t, a given peak i_0 is associated to the corresponding nearest pitch number p_0 and $A(p_0,t)$ is set to $A(p_0,t) = P_h(i_0-1,t)+P_h(i_0,t)+P_h(i_0+1,t)$. A is then normalized as follows : $A \leftarrow A/\max_{p,t} A(p,t)$.
- Finally, onset/offset detection is employed to transcribe note events. For each pitch p, an onset (resp. an offset) is detected as soon as $A(p,t)$ becomes larger (resp. lower) than $A_{\min}$ for more than 70 ms, $A_{\min}$ being a fixed threshold. In order to consider onsets of repeating notes (i.e. when the energy of pitch p remains above the threshold $A_{\min}$ whereas there is a new onset), we applied another onset detection on the derivative $A'(p,t)$ of $A(p,t)$ w.r.t. time. A new onset is detected once $A'(p,t) > A'_{\min}$, $A'_{\min}$ being another fixed threshold. If two detected onsets of a same MIDI note are closer than 100 ms, only the first one is kept.

2) Metrics: once for all A given estimated note is considered to be correctly transcribed if a note with a same pitch and onset time (within 50 ms) can be found in the ground truth. Traditional measures of recall $\mathcal{R}$, precision $\mathcal{P}$ and F-measure $\mathcal{F}$ [34] are then calculated to assess the performance.

3) Playing With the Parameters on a Training Database: Five 30 s excerpts from the RWC classical genre database [35], all listed in Table II and referred to as database DB_{train}, were used for those preliminary evaluations.

As already mentioned in Section III-A-1, in practice, one source represents a "meta-instrument" rather than a real instrument. Consequently, there is no need to set the number of sources S to the actual number of instruments in the input signal: a fixed number of sources can be sufficient to model an unknown number of instruments. In order to both verify this statement and to find a good value for S (so that we can fix it once for all), in a first experiment, the proposed transcription algorithm has been performed using different values of S on

[1]http://audition.ens.fr/adc/

TABLE III
STUDY OF THE INFLUENCE OF THE VALUE OF S: F-MEASURE FOR EACH FILE OF DB_{train} WHEN ONSET THRESHOLDS ARE OPTIMALLY SET

S	rwc(1)	rwc(2)	rwc(3)	rwc(4)	rwc(5)	Mean
1	72.2	31.1	43.3	43.5	73.8	47.0
2	76.7	32.0	45.7	45.1	80.3	50.0
3	78.9	**33.3**	**46.0**	45.4	80.7	50.5
4	81.4	33.1	45.7	**45.8**	81.5	**51.1**
5	**82.0**	32.3	46.0	45.3	**82.2**	50.7
6	82.0	32.7	47.1	45.2	82.2	51.0

each file of DB_{train}. In this experiment, no prior was added and the onset detection thresholds A_{min} and A'_{min} have been optimally set for each file and each value of S. Results are reported in Table III. For each file, by comparing the value of S which maximizes the performance, and the effective number of instruments (see Table II), one can observe that there is no correlation between those two quantities. Indeed, best results are always given for S between 3 and 5, regardless of the number of instruments. We can then conclude that setting S to a fixed value can be sufficient to model any number of instruments and that it will not have a negative effect onto the transcription performance. From now on, S is fixed to 4, which is the best value in this first experiment, in terms of mean results.

The aim of the second test is to evaluate the influence of the use of the different priors on the transcription system. To do so, four different systems, all described in Table IV, have been evaluated on this database: H_4 will refer to the basic model with no prior, $H_4 - t$ to the system with the temporal prior alone, $H_4 - s$ to the system with the sparse prior alone, and finally $H_4 - st$ with both priors. The subscript 4 means that the number of sources is $S = 4$. From experiments that we performed on this same database, hyperparameters β and χ have been manually set to optimal values, as well as the onset threshold A'_{min}. Results are showed in Fig. 7, where the average F-measure w.r.t. the onset threshold A_{min} is plot for each system. A first remark is that the system seems to behave similarly whether the spectral envelope temporal continuity prior is active or not, even though this prior does affect the estimation of the parameters (see for instance Fig. 5). One explanation could rely in the post-processing stage for the automatic transcription (the calculation of the resulting pitch activity matrix $A(p,t)$ as well as the onset/offset detection) which seems to smooth the lack of temporal continuity when the prior is not used. We can also wonder why the continuity prior did have a clear positive impact on the monopitch evaluation. It appears that the monomodal prior makes the algorithm more sensitive to local maxima (basically it can lead to octave errors). The role of the continuity prior is then to avoid those local minima. When no monomodal prior is used, as in this experiment, the role of the continuity prior is thus weakened. It can be noticed though that on this database, it slightly improves the maximum F-measure, when combined to the sparse prior. Concerning this last prior, its effect can be clearly seen: besides increasing the maximal value of $\mathcal{F}$, it allows the system being a lot less sensitive to a suboptimal value of A_{min}. Since the optimal threshold might change according to the file or the database, this characteristic is quite interesting. From those preliminary results it is possible to determine good values of the fixed parameters for each version of the HALCA algorithm. Those values are summarized in Table V. It can be noticed that the chosen thresholds A_{min} are slightly lower than the ones that maximize $\mathcal{F}$ on DB_{train}. The reason is that the F-measure curves decrease faster if A_{min} is overestimated than if it is underestimated. Doing so, we hope that the proposed algorithms will be more robust to the nature of the evaluation dataset. Only the three systems H, $H_4 - s$ and $H_4 - st$ are kept for comparison with other transcription algorithms from the literature.

TABLE IV
PROPOSED AND REFERENCE ALGORITHMS

Symbol	Description
H_4	HALCA model with no prior.
$H_4 - t$	HALCA model with spectral envelope temporal continuity prior.
$H_4 - s$	HALCA model with sparseness prior.
$H_4 - st$	HALCA model with spectral envelope temporal continuity and sparseness priors.
Vincent'10 [13]	Multiplicative NMF with the β-divergence ($\beta = 0.5$) and harmonicity and spectral smoothness constraints.
Dessein'12 [10]	Spectrogram decomposition on a learned dictionary using β-divergence. The dictionary is learned on isolated notes of piano (from MAPS database [36]).

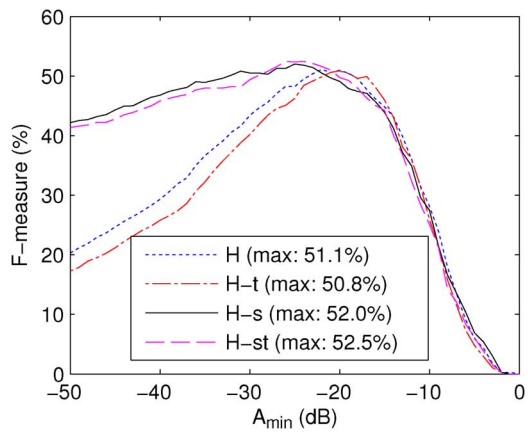


Fig. 7. Study of the influence of the different priors: average F-measure w.r.t. A_{min} for four versions of the proposed transcription system.

TABLE V
VALUE OF THE FIXED PARAMETERS FOR THE PROPOSED ALGORITHMS

System	S	$A_{min}(dB)$	A'_{min}	β	χ
H_4	4	-25	0.018	0	0
$H_4 - s$	4	-30	0.018	0.06	0
$H_4 - st$	4	-30	0.018	0.06	10^7

TABLE VI
DB_{eval}: LIST OF THE AUDIO EXCERPTS FROM MIREX 2007 MULTI-F0 DEVELOPMENT DATASET [37] AND QUASI TRANSCRIPTION CORPUS [38]

Symbol	Title (Artist)	Duration	Genre	Polyphony level (Ave. / Max.)
mir	Woodwind Quintet (Beethoven)	$0'54''$	Classical	3.1 / 6
qua(1)	One we love (Another Dream)	$3'25''$	Pop	1.8 / 4
qua(2)	The Spirit of Shackleton (Glen Philips)	$4'04''$	Alternative	3.9 / 15
qua(3)	Mix Tape (Jims Big Ego)	$3'03''$	Rock	3.8 / 8
qua(4)	Good Soldier (Nine Inch Nails)	$3'22''$	Industrial Rock	3.8 / 10
qua(5)	The Ultimate NZ Tour (?)	$2'21''$	Pop	5.0 / 12
qua(6)	Ana (Vieux Farka Toure)	$4'09''$	Reggae	1.7 / 6

TABLE VII
TRANSCRIPTION RESULTS: F-MEASURE (%) FOR EACH FILE OF DB_{eval}

File	H_4	$H_4 - s$	$H_4 - st$	Vincent'10	Dessein'12
mir	**62.0**	58.8	60.8	57.9	52.0
qua(1)	34.3	40.8	**41.8**	10.0	19.0
qua(2)	15.8	**16.7**	15.6	8.4	7.8
qua(3)	**11.1**	9.5	9.0	6.8	2.5
qua(4)	19.6	**27.6**	27.0	8.0	13.6
qua(5)	18.3	16.6	**18.5**	9.8	13.7
qua(6)	**47.9**	46.9	46.4	9.5	4.3
mean	29.9	31.0	**31.3**	15.8	16.1

TABLE VIII
MEAN RESULTS AND COMPUTATION TIME (CT) ON DB_{eval}

Algorithm	$\mathcal{F}$ (%)	$\mathcal{R}$ (%)	$\mathcal{P}$ (%)	CT ($\times$ real time)
H_4	29.9	27.9	37.0	3.4
$H_4 - s$	31.0	26.6	**40.3**	4.3
$H_4 - st$	**31.3**	27.6	38.6	7.5
Vincent'10	15.8	**48.0**	10.6	0.9
Dessein'12	16.1	20.1	14.9	0.8

4) Comparing With Other Methods on an Evaluation Database: In order to compare the performance of the two systems $H_4 - s$ and $H_4 - st$ to other state-of-the-art algorithms, a second database denoted DB_{eval} and composed of 7 audio files has been set up. The first file of DB_{eval} is a 54 s excerpt from a woodwind quintet transcription of Beethoven's string quartet No.4 Op.18, available in the MIREX 2007 multi-F0 development dataset [37]. The other recordings are from the QUASI transcription corpus [38]. Each file is briefly described in Table VI.

The performance of the algorithms $H_4 - s$ and $H_4 - st$ are compared with two other NMF based transcription systems, namely Vincent'10 [13] and Dessein'12 [10] all described in Table IV. Those systems are run from their author's implementation, which they kindly shared. Global transcription results ($\mathcal{F}$) for each file of DB_{eval} are reported in Table VII whereas mean transcription metrics and computation time[2] are reported in Table VIII. From those tables, several interesting results can be pointed out. First, the same conclusion as in Section V-C-3 can be drawn: results from the two systems $H_4 - s$ and $H_4 - st$ are very similar and the addition of the temporal prior does not significantly improve the performances of the transcription system. Concerning the use of the sparse prior, it can be seen that it slightly improves the global F-measure. In any case, the algorithms we provide have the best F-measure for each file of DB_{eval}, and therefore the best average F-measure. One interesting characteristic can be deduced from Table VII: the relative difference between the performance of H, $H_4 - s$ or $H_4 - st$ and the performance of Vincent'10 or Dessein'12 is much greater for the files of QUASI corpus (qua(n)) than for the Mirex woodwind quartet (mir). This highlights the relative robustness of the proposed model to the musical genre. The conclusion one can draw from Table VIII is that each transcription system would benefit from including an automatic estimation of the threshold of note detection (A_{min} in our case). In fact, the difference between $\mathcal{R}$ and $\mathcal{P}$ can teach us that this threshold is not optimal for each file of DB_{eval}, whereas it has been set so that it is optimal on some other dataset (DB_{train} in our case).

VI. CONCLUSION

In this paper, we proposed a new PLCA-based model called HALCA which analyzes harmonic structures in musical signals. The model does not rely on an hypothesis of redundancy, and therefore is quite expressive. Particularly, it allows modeling sources that have both temporal variations of pitch and spectral envelope. The presence of noise is also considered in order to enforce the robustness of the model to real data. Three new priors that help the algorithm to converge toward a meaningful solution have also been proposed. Those priors are generic and can be applied to any other PLCA-based model. The HALCA model has first been tested on a task of monopitch estimation and the results have showed that it could fit to many kinds of musical instruments. Then, it has been tested on a task of automatic transcription on two different databases. On the training one, the effect of using the different priors has been studied. It appeared that the use of the sparse prior has an impact on the sensitivity to the onset detection threshold. On the contrary, the

[2]Evaluations have been done using a 64-bit version of Matlab on a 3.1 GHz processor.

use of the timbre temporal continuity does not seem to significantly improve the results in this task of automatic transcription. The training database is also used to fix the hyperparameters of the system. On the second database, three versions of the HALCA algorithm are compared to two state-of-the-art algorithms. The results our systems provide and their comparison with the reference systems permit to outline the following conclusions. First, even if the addition of a sparse prior can improve the performance of a threshold-based onset detection, we think that every system would benefit from being able to automatically estimate this threshold w.r.t. the input signal. In future work, we plan to imagine a model where the sparse prior is replaced by the estimation of the best onset detection threshold. The second conclusion is that maybe the hypothesis of redundancy is not necessary for a TFR factorization technique to be efficient, and a highly expressive model is recommended to be robust to any kind of real musical signal.

The Matlab implementation of the algorithm can be found online : http://www.tsi.telecom-paristech.fr/aao/en/2012/07/13/fuentes2012_ieee/.

Appendix A
EM Update Rules With Sparse Prior

One want to find the argument $\hat{\boldsymbol{\theta}}$ of the maximum of the function S defined in (24) under the constraint $\varphi(\boldsymbol{\theta}) = 1 - \sum_j \theta_j = 0$. We know that the maximum exists since S is bounded on Ω and that it verifies first and second order necessary conditions, proper to local minima (S and φ are twice differentiable): there exists an unique $\rho \in \mathbb{R}$ such that ($\langle .,. \rangle$ denotes the scalar product)

$$\nabla L_\rho(\hat{\boldsymbol{\theta}}) = 0 \text{ and} \tag{29}$$

$$\left\langle H\left(L_\rho(\hat{\boldsymbol{\theta}})\right) d, d \right\rangle \leq 0, d \in \left\{ d \in \mathbb{R}^J / \langle \nabla \varphi(\hat{\boldsymbol{\theta}}), d \rangle = 0 \right\}, \tag{30}$$

where L_ρ is the Lagrange function defined as:

$$\begin{aligned} L_\rho : \Omega &\longrightarrow \mathbb{R} \\ \boldsymbol{\theta} &\longmapsto \mathcal{S}(\boldsymbol{\theta}) + \rho\, \varphi(\boldsymbol{\theta}), \end{aligned} \tag{31}$$

and where $H(L_\rho(\hat{\boldsymbol{\theta}}))$ is the Hessian matrix of L_ρ at point $\hat{\boldsymbol{\theta}}$. Equation (29) leads to:

$$\forall j, \frac{w_j}{\hat{\theta}_j} - \frac{\beta\sqrt{J}}{\sqrt{\hat{\theta}_j}} - \rho = 0. \tag{32}$$

If $\rho = 0$, then

$$\forall j, \hat{\theta}_j = \frac{w_j^2}{\beta^2 J} \tag{33}$$

which is possible only if $\sum_j w_j^2/(\beta^2 J) = 1$.

If $\rho > 0$, then

$$\forall j, \sqrt{\hat{\theta}_j} = \frac{-\beta\sqrt{J} + \sqrt{\beta^2 J + 4\rho w_i}}{2\rho}$$

that is to say:

$$\forall j, \hat{\theta}_j = \frac{2w_j^2}{J\beta^2 + 2\rho w_j + \beta\sqrt{J}\sqrt{\beta^2 J + 4\rho w_j}}, \tag{34}$$

which is possible only if $\sum_j w_j^2/(\beta^2 J) > 1$. In this case, we can prove that there is a unique $\rho > 0$ such that $\sum_j \hat{\theta}_j = 1$.

If $\max_j -\beta^2 J/4w_j \leq \rho < 0$, then,

$$\forall j, \hat{\theta}_j = \frac{2w_j^2}{J\beta^2 + 2\rho w_j \pm \beta\sqrt{J}\sqrt{\beta^2 J + 4\rho w_j}}, \tag{35}$$

which leads to 2^J possible solutions. However, if we take a vector d such that

$$\begin{cases} d_{j_1} = 1, \quad j_1 \in [[1, J]] \\ d_{j_2} = -1, \quad j_2 \in [[1, J]] \setminus \{j_1\} \\ d_j = 0, \quad \forall j \in [[1, J]] \setminus \{j_1, j_2\} \end{cases},$$

condition (30) leads to the following condition:

$$-\frac{w_{j_1}}{\hat{\theta}_{j_1}^2} + \frac{\beta}{2\hat{\theta}_{j_1}^{\frac{3}{2}}} - \frac{w_{j_2}}{\hat{\theta}_{j_2}^2} + \frac{\beta}{2\hat{\theta}_{j_2}^{\frac{3}{2}}} < 0. \tag{36}$$

It can then be deduced that, $\exists j_0 / \forall j \neq j_0, \hat{\theta}_j < 4w_j^2/(\beta^2 J)$ and therefore, $\exists j_0$ such that:

$$\forall j \neq j_0, \hat{\theta}_j = \frac{2w_j^2}{J\beta^2 + 2\rho w_j + \beta\sqrt{J}\sqrt{\beta^2 J + 4\rho w_j}}. \tag{37}$$

We have then reduced the number of possible solutions for $\hat{\boldsymbol{\theta}}$. In order to find the global maximum, it is possible to check which solution maximizes function S.

Nevertheless, in practice, we set β to a sufficiently low value so that the statement $\sum_j w_j^2/(\beta^2 J) > 1$ is always true. In this case, $\rho > 0$ is the only possible solution and $\hat{\boldsymbol{\theta}}$ is given by (34). ρ can be found with any numerical root finding algorithm. Fig. 4 illustrates the advantage of using such a prior.

References

[1] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, ser. Synthesis Lectures on Speech and Audio Processing. San Rafael, CA, USA: Morgan & Claypool, 2009.

[2] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, USA, October 2003, pp. 177–180.

[3] 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.

[4] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard, "Probabilistic model for main melody extraction using constant-Q transform," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 5357–5360.

[5] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

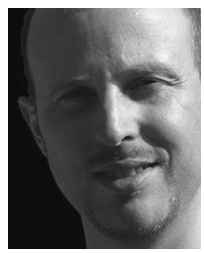
[6] K. Ochiai, H. Kameoka, and S. Sagayama, "Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 133–136.

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

[8] G. Mysore and P. Smaragdis, "Relative pitch estimation of multiple instruments," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 313–316.

[9] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non stationary audio events," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 744–753, May 2011.
[10] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, F. Nielsen and R. Bhatia, Eds. New York, NY, USA: Springer, July 2012.
[11] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a convolutive probabilistic model," in *Proc. SMC*, Padova, Italy, Jul. 2011, pp. 19–24.
[12] G. Grindlay and D. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1159–1169, Oct. 2011.
[13] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.
[14] S. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. ISMIR*, Vienna, Austria, Sep. 2007, pp. 381–386.
[15] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proc. ICASSP*, Las Vegas, NV, USA, April 2008, pp. 2069–2072.
[16] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–549, Mar. 2010.
[17] B. Fuentes, R. Badeau, and G. Richard, "Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA," in *Proc. ICASSP*, Prague, Czech Republic, May 2011, pp. 401–404.
[18] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. EUSIPCO*, Bucharest, Romania, Aug. 2012, pp. 2654–2658.
[19] D. Fitzgerald, M. Cranitch, and E. Coyle, "Shifted 2D non-negative tensor factorisation," in *Proc. ISSC*, Dublin, Ireland, 2006.
[20] H. Kameoka, T. Nishimoto, and S. S. , "A multipitch analyser based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.
[21] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.
[22] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 1825–1828.
[23] M. Shashanka, "Latent variable framework for modeling and separating single-channel acoustic sources," Ph.D. dissertation, Boston Univ., Boston, MA, USA, Aug. 2007.
[24] M. N. Schmidt and H. Laurberg, "Non-negative matrix factorization with Gaussian process priors," *Comput. Intell. Neurosci.*, pp. 1–10, 2008.
[25] K. Y. Yilmaz, A. T. Cemgil, and U. Simsekli, "Generalized coupled tensor factorization," in *NIPS*, Granada, Spain, Dec. 2011.
[26] M. Nakano, J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Infinite-state spectrum model for music signal analysis," in *Proc. ICASSP*, Prague, Czech Republic, May 2011, pp. 1972–1975.
[27] J. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.
[28] J. Prado, Une Inversion Simple de la Transformée à Q Constant, [Online]. Available: http://www.tsi.telecom-paristech.fr/aao/en/2011/06/06/inversible-cqt/ 2011
[29] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, Ed. Berkeley, CA, USA: Univ. of California Press, 1951, pp. 481–492.
[30] W. Mekwi, "Iterative methods for roots of polynomials," M.S. thesis, Univ. of Oxford, Oxford, U.K., 2001.
[31] G. Grindlay and D. Ellis, "A probabilistic subspace model for multi-instrument polyphonic transcription," in *Proc. ISMIR*, Utrecht, The Netherlands, Aug. 2010, pp. 21–26.
[32] University of Iowa Musical Instrument Sample Database [Online]. Available: http://theremin.music.uiowa.edu/index.html
[33] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.
[34] C. van Rijsbergen, *Information Retrieval*, 2nd ed. London, U.K.: Butterworths, 1979.
[35] M. Goto and T. Nishimura, "Rwc music database: Music genre database and musical instrument sound database," in *Proc. ISMIR*, Oct. 2003.
[36] V. Emiya, "Transcription automatique de la musique de piano," Ph.D. dissertation, Telecom ParisTech, Paris, France, 2008.
[37] Music Information Retrieval Evaluation Exchange (MIREX), [Online]. Available: http://music-ir.org/mirexwiki/
[38] QUASI-Transcription, Set 1, v1.0, [Online]. Available: http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

Benoit Fuentes was born in France in 1985. He received the State Engineering degree from Télécom ParisTech, Paris, France, in 2008, 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 2009, and the Ph.D. degree in signal processing from Télécom ParisTech, Paris, in 2013. His research focuses on probabilistic modeling of audio signals applied to automatic transcription and source separation.



Roland Badeau (M'02–SM'10) was born in Marseille, France, in 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. He received the ParisTech Ph.D. Award in 2006, and the Habilitation degree from the Université Pierre et Marie Curie (UPMC), Paris VI, in 2010. In 2001, he joined the Department of Signal and Image Processing of Télécom ParisTech, CNRS LTCI, as an Assistant Professor, where he became Associate Professor in 2005. From November 2006 to February 2010, he was the manager of the DESAM project, funded by the French National Research Agency (ANR), whose consortium was composed of four academic partners. His research interests focus on statistical modeling of non-stationary signals (including adaptive high resolution spectral analysis and Bayesian extensions to NMF), with applications to audio and music (source separation, multipitch estimation, automatic music transcription, audio coding, audio inpainting). He is a co-author of 21 journal papers, over 50 international conference papers, and 2 patents. He teaches in the Master of Engineering of Télécom ParisTech and in the Master of Sciences and Technologies of UPMC. He is also a Chief Engineer of the French Corps of Mines (foremost of the great technical corps of the French state) and an Associate Editor of the EURASIP Journal on Audio, Speech, and Music Processing.

Gaël Richard (SM'06) received the State Engineering degree from Télécom ParisTech, France (formerly ENST) in 1990, the Ph.D. degree from LIMSI-CNRS, University of Paris-XI, in 1994 in speech synthesis, and the Habilitation Ã Diriger des Recherches degree from the University of Paris XI in September 2001. After the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. From 1997 to 2001, he successively worked for Matra, Bois d'Arcy, France, and for Philips, Montrouge, France. In particular, he was the Project Manager of several large scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, Telecom ParisTech, where he is now a Full Professor in audio signal processing and Head of the Audio, Acoustics, and Waves research group. He is a coauthor of over 120 papers and inventor in a number of patents and is also one of the experts of the European commission in the field of audio signal processing and man/machine interfaces. He was an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING between 1997 and 2011 and one of the guest editors of the special issue on "Music Signal Processing" of IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING (2011). He currently is a member of the IEEE AUDIO AND ACOUSTIC SIGNAL PROCESSING TECHNICAL COMMITTEE, member of the EURASIP and AES and senior member of the IEEE.