

Adaptive harmonic spectral decomposition for multiple pitch estimation

Emmanuel Vincent, Nancy Bertin, Roland Badeau

▶ **To cite this version:**

Emmanuel Vincent, Nancy Bertin, Roland Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. [Research Report] PI 1919, 2009, pp.15. inria-00350163v2

HAL Id: inria-00350163

https://inria.hal.science/inria-00350163v2

Submitted on 7 Jul 2009 (v2), last revised 10 Dec 2010 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

I R I S A
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION
INTERNE
N° 1919

ADAPTIVE HARMONIC SPECTRAL DECOMPOSITION FOR MULTIPLE PITCH ESTIMATION

EMMANUEL VINCENT, NANCY BERTIN AND ROLAND BADEAU

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

Adaptive harmonic spectral decomposition for multiple pitch estimation

Emmanuel Vincent[*] , Nancy Bertin[**] and Roland Badeau[***]

Systèmes cognitifs
Projet METISS

Publication interne n ° 1919 — Janvier 2009 — 15 pages

Abstract: Multiple pitch estimation consists of inferring the fundamental frequencies and the salience of the notes forming a music signal over short time frames. This mid-level representation can be exploited as a front-end for higher-level applications, such as music-to-score transcription or chord detection. One approach is to decompose the short-term magnitude spectrum of the signal into a sum of basis spectra representing individual pitches scaled by time-varying amplitudes, using algorithms such as nonnegative matrix factorization (NMF). Prior training of the basis spectra is often infeasible due to the wide range of possible instruments. Appropriate spectra must then be estimated from the observed data, which may result in limited performance due to inaccurately estimated spectra. In this article, we model each basis spectrum as a weighted sum of narrowband spectra representing a few adjacent harmonic partials, thus enforcing harmonicity and spectral smoothness while adapting the spectral envelope to each instrument. We derive a NMF-like algorithm to estimate the model parameters and evaluate it on a database of piano recordings, considering several choices for the narrowband spectra. Performance appears superior to unconstrained adaptive NMF and competitive with supervised NMF based on pre-trained piano spectra. We also apply our approach to woodwind data.

Key-words: Multiple pitch estimation, adaptive representation, nonnegative matrix factorization, harmonicity, spectral smoothness

(Résumé : tsvp)

N. Bertin and R. Badeau are supported by ANR young researchers project DESAM.

[*] METISS group, IRISA-INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France – emmanuel.vincent@irisa.fr
[**] Institut Télécom, Télécom ParisTech, 37-39 rue Dareau, 75014 Paris, France – nancy.bertin@telecom-paristech.fr
[***] Institut Télécom, Télécom ParisTech/CNRS-LTCI, 37-39 rue Dareau, 75014 Paris, France – roland.badeau@telecom-paristech.fr

Centre National de la Recherche Scientifique
(UMR 6074) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

Décomposition spectrale harmonique adaptative pour l'estimation de fréquences fondamentales multiples

Résumé : L'estimation de hauteurs multiples consiste à inférer la fréquence fondamentale et la saillance des notes formant un signal de musique sur des trames de courte durée. Cette représentation de moyen niveau peut être exploitée en tant que représentation frontale pour des applications de plus haut niveau, comme la transcription sous forme de partition musicale ou la détection d'accords. Une approche possible est de décomposer le spectre d'amplitude à court terme du signal en une somme de spectres de base représentant chacun une hauteur multipliés par une amplitude variant au cours du temps, à l'aide d'algorithmes tels que la factorisation matricielle positive (*nonnegative matrix factorization* ou NMF). L'apprentissage préalable des spectres de base est souvent irréalisable à cause du grand nombre d'instruments possibles. Des spectres appropriés doivent alors être estimées à partir des données observées, avec le risque que les erreurs d'estimation se traduisent par des erreurs sur les hauteurs. Dans cet article, nous modélisons chaque spectre de base comme une somme pondérée de spectres à bande étroite représentant quelques partiels harmoniques adjacents, de façon à garantir l'harmonicité et la lissité des spectres tout en adaptant leur enveloppe spectrale à chaque instrument. Nous proposons un algorithme de type NMF pour estimer les paramètres du modèle et nous l'évaluons sur un ensemble d'enregistrements de piano, en considérant différents choix de spectres à bande étroite. La performance apparaît supérieure à celle de la NMF sans contrainte et compétitive avec celle de la NMF supervisée reposant sur des spectres appris préalablement. Nous appliquons aussi notre approche à des enregistrements d'instruments à vent.

Mots clés : Estimation de fréquences fondamentales multiples, représentation adaptative, factorisation matricielle positive, harmonicité, lissité spectrale

1 Introduction

Music signals involve a collection of sounds, which may be either pitched or unpitched. Multiple pitch estimation consists of inferring the fundamental frequencies of pitched sounds over short time frames and quantifying confidence in these estimates by means of a salience measure [1]. This mid-level representation is often processed via hidden Markov modeling to transcribe a signal into higher-level note events characterized by their onset time, duration, pitch and voice [2]. It can also be exploited as a front-end for many other applications, including chord detection [3], instrument identification [4] and source separation [5].

A variety of approaches have been proposed to address multiple pitch estimation in the literature [1], ranging from correlograms [6] and harmonic sum [7] to neural networks [8], time-domain generative models [9, 10] and support vector machines [11]. One particular approach is to decompose the short-term magnitude or power spectrum of the signal into a sum of basis spectra representing individual pitches scaled by time-varying amplitudes. The basis spectra can be either fixed by training on annotated recordings [12, 13, 14] or adaptively estimated from the observed spectra [15, 16, 17, 18, 19]. The parameters of this model can be estimated by nonnegative matrix factorization (NMF), sparse decomposition or sparse dictionary learning. These algorithms minimize distortion between observed and model spectra, given some optional temporal priors such as continuity and sparsity. Fixed basis spectra typically achieve better performance, provided that test and training data involve the same instruments in similar recording conditions, which is difficult to satisfy in practice. Adaptive basis spectra address this issue, but result in limited performance due to the lack of constraints ensuring that each basis spectrum has a clearly identifiable pitch. Constraints of spectral shift invariance [20] or source-filter modeling [21] favor more structured spectra. However they do not guarantee that the estimated spectra are harmonic. Experiments in [22] suggest that these constraints are respectively inappropriate and insufficient: shift invariance does not account for variations of spectral envelope as a function of pitch, while source-filter modeling includes a large number of parameters that are difficult to estimate reliably.

A more principled approach to the estimation of adaptive pitched basis spectra is to design explicit harmonicity constraints. In [23], each basis spectrum is constrained to zero in all bins but the multiples of a fixed fundamental frequency. This model relies on a crude approximation of the spectrum of a sinusoidal partial and is prone to errors since the harmonicity constraint alone does not allow segregation between a given fundamental frequency and its submultiples. In [24, 22], each basis spectrum is modeled as a weighted sum of spectra representing individual partials and the weights are constrained via a source-filter model, where the source weights are either trained specifically for singing voice [24] or estimated from the test data [22]. This additional constraint appears efficient in the context of melody transcription or source separation, provided each instrument plays a sufficient number of different pitches and its observed pitch range is known [22]. In [25, 26], we introduced a different approach whereby each basis spectrum is modeled as a weighted sum of narrowband spectra with a smooth envelope representing a few adjacent harmonic partials. This approach reduces octave errors without assuming prior dependencies between the spectral envelopes of different pitches. It is perhaps closer to low-level auditory processing of pitch, which relies on the presence of several partials within certain auditory bands [1]. Inharmonicity and variable tuning constraints were also explored in [26] but did not bring any improvement.

In this article, we further investigate the use of harmonicity and spectral smoothness as explicit constraints for NMF-based adaptive spectral decomposition, independently of any temporal prior. We extend our preliminary work in several ways. Firstly, we study several definitions for the narrowband spectra, including training from annotated recordings. Secondly, we consider a range of distortion measures. Thirdly, we evaluate our algorithm on a more diverse database, compare it to the alternative approaches discussed above and quantify its robustness to the chosen parameter values. The structure of the rest of the article is as follows. In Section 2, we describe baseline NMF-based algorithms and provide example results. We present the proposed adaptive harmonic model and the associated algorithm in Section 3. We evaluate these algorithms on a database of music recordings in Section 4 and conclude in Section 5.

2 Baseline decompositions over fixed or unconstrained basis spectra

Baseline NMF-based algorithms for multiple pitch estimation involve the following steps: computing a time-frequency representation of the signal, decomposing it into a scaled sum of fixed or adaptive basis spectra, identifying the pitch of each spectrum in the latter case and deriving a pitch salience measure from the associated time-varying amplitudes. Each of these steps involves some design choices outlined below.

2.1 ERB-scale time-frequency representation

In order to discriminate musical pitches, the time-frequency representation must have a resolution of at least one semitone over the whole frequency range. This can be achieved using the short-time Fourier transform (STFT) with a long window [17], a constant-Q filterbank [20] or another nonuniform filterbank. In the following, we consider the auditory-motivated filterbank in [13]. The input signal is passed through a set of $F = 250$ filters indexed by f consisting of sinusoidally modulated Hann windows with frequencies ν_f linearly spaced between 5 Hz and 10.8 kHz on the Equivalent Rectangular Bandwidth (ERB) scale [27] given by $\nu_f^{\mathrm{ERB}} = 9.26 \log(0.00437\nu_f^{\mathrm{Hz}} + 1)$. The length L_f of each filter is set so that the bandwidth of its main frequency lobe equals four times the difference between its frequency and those of adjacent filters. Each subband is then partitioned into disjoint 23 ms time frames indexed by t and and the root-mean-square magnitude X_{ft} is computed within each frame. This yields similar pitch estimation performance to the STFT at a lower computation cost due to reduction of the number of frequency bins [25].

2.2 Magnitude-domain NMF with β-divergence

NMF refers to a set of algorithms minimizing some distortion measure between the observed spectrum X_{ft} and the model spectrum Y_{ft} defined as

$$Y_{ft} = \sum_{i=1}^{I} A_{it} S_{if} \tag{1}$$

where S_{if} and A_{it}, $1 \leq i \leq I$, are a set of basis spectra and time-varying amplitudes, respectively. This model has been applied to magnitude spectra [15] or, more rarely, to power spectra [13]. Different parametric distortion measures have been employed within the family of β-divergences [28]

$$d(X_{ft}|Y_{ft}) = \frac{1}{\beta(\beta-1)}(X_{ft}^{\beta} + (\beta-1)Y_{ft}^{\beta} - \beta X_{ft} Y_{ft}^{\beta-1}), \tag{2}$$

including the Euclidean distance ($\beta = 2$) [15], Kullback-Leibler divergence ($\beta \to 1$) [15] and Itakura-Saito divergence ($\beta \to 0$) [16], or within the family of perceptually weighted Euclidean distances [25]. Both families involve a parameter $\beta \geq 0$ that can be chosen so that the distortion scales with X_{ft}^{β}. A small β compresses the large dynamic range of music, hence increasing the modeling accuracy of quiet sounds.

The model parameters can be estimated either by inferring both adaptive basis spectra and time-varying amplitudes from the test data or by learning fixed basis spectra from training data and inferring their time-varying amplitudes only from the test data. Training and inference are both achieved by minimization of the chosen distortion measure. After suitable initialization of the parameters, the β-divergence can be minimized by iterative application of one or both of the following multiplicative updates rules until convergence [28]

$$A_{it} \leftarrow A_{it} \frac{\sum_{f=1}^{F} S_{if} Y_{ft}^{\beta-2} X_{ft}}{\sum_{f=1}^{F} S_{if} Y_{ft}^{\beta-1}} \tag{3}$$

$$S_{if} \leftarrow S_{if} \frac{\sum_{t=1}^{T} A_{it} Y_{ft}^{\beta-2} X_{ft}}{\sum_{t=1}^{T} A_{it} Y_{ft}^{\beta-1}}. \tag{4}$$

In the following, initialization is achieved either by randomly drawing A_{it} and S_{if} from a uniform distribution when estimating the spectra or by setting A_{it} to 1 when using fixed spectra. Although it has been proved that β-divergence is nonincreasing under these update rules for $1 \leq \beta \leq 2$ only [29], experimental convergence has been observed for any β [28, 19]. Similar updates can be derived for weighted Euclidean distances [25].

We compared the aforementioned modeling domains and distortion measures using the data and the F-measure metric defined in Section 4. Similar figures were obtained with adaptive basis spectra. However, with fixed pre-trained spectra, the average F-measure decreased by 8% with power-domain modeling instead of magnitude-domain modeling and by 11% with the perceptually weighted Euclidean distance instead of β-divergence. Therefore, despite the fact that power-domain modeling better approximates linear combination of note signals in the time domain, we focus on magnitude-domain NMF with β-divergence in the following.

2.3 Harmonic comb-based pitch identification

The pitch p_i of a given basis spectrum S_{if} is measured on the Musical Instrument Digital Interface (MIDI) semitone scale and is related to its fundamental frequency ν_{i0}^{Hz} via

$$\nu_{i0}^{\mathrm{Hz}} = 440 \times 2^{\frac{p_i - 69}{12}}. \tag{5}$$

When training the basis spectra on annotated data, each basis spectrum is associated a priori with a fixed integer pitch and accurate training is ensured by setting to zero the amplitudes of the basis spectra corresponding to inactive pitches. By contrast, basis spectra estimated from the test data may be either pitched or unpitched and their pitches must be found a posteriori. In the following, we use the sinusoidal comb estimator [25]

$$\nu_{i0}^{\mathrm{Hz}} = \arg\min_{\nu_0^{\mathrm{Hz}}} \sum_{f=1}^{F} S_{if}^2 \left[1 - \cos(2\pi \nu_f^{\mathrm{Hz}} / \nu_0^{\mathrm{Hz}})\right]. \tag{6}$$

The pitch range is chosen as the interval between $p_{\mathrm{low}} = 21$ (27.5 Hz) and $p_{\mathrm{high}} = 108$ (4.19 kHz), which is the range of the piano. The basis spectra whose estimated pitch is outside this range are classified as unpitched. We found that, despite its simplicity, this estimator was surprisingly efficient for the post-processing of basis spectra estimated via NMF. Alternative spectral-domain pitch estimators, such as the popular YinFFT [30], did not perform as well in this context.

2.4 Amplitude-based pitch salience measure

Given the time-varying amplitudes of all basis spectra, we measure the salience of an integer pitch p by the square root of the total power of the scaled basis spectra whose pitch p_i is within one quarter-tone of p

$$\bar{A}_{pt} = \left(\sum_{f=1}^{F} \left(\sum_{i \text{ s.t. } |p_i - p| < 1/2} A_{it} S_{if} \right)^2 \right)^{1/2}. \tag{7}$$

This measure scales as an amplitude and is hence comparable to other amplitude-based measures, such as the harmonic sum in [7]. Due to their real-valued output, such measures cannot be directly compared to ground truth annotations which characterize a given pitch as either active or inactive. Instead, we derive pitch estimates on a frame-by-frame basis by classifying a given pitch p as active whenever

$$\bar{A}_{pt} \geq 10^{-A_{\min}/20} \max_{pt} \bar{A}_{pt} \tag{8}$$

where $A_{\min}$ is a detection threshold in decibels (dB) that can be either set manually or learned from training data. We found that this decision strategy was more efficient than the one in [7] for the estimation of the number of active pitches per frame.

2.5 Example results

Figs. 1 to 3 illustrate the multiple pitch estimation results derived from NMF with adaptive or fixed basis spectra over an excerpt of Borodin's *Little Suite - Serenade*, recorded from an acoustic piano and taken from the MIDI-Aligned Piano Sounds (MAPS) database [31]. The number of basis spectra was set to $I = p_{\mathrm{high}} - p_{\mathrm{low}} + 1 = 88$ and β was set to its optimal value determined in Section 4. Training was conducted on the University of Iowa's musical instrument samples (MIS) [32], which include isolated note sounds from a single piano at all pitches and at three loudness levels. The detection threshold $A_{\min}$ was set to 25 dB.

We observe that many basis spectra estimated via adaptive NMF are neither clearly pitched nor unpitched. Most spectra involve spurious spectral peaks besides the predominant harmonic series or missing peaks in that series. Some spectra even represent several pitches at a time. The resulting pitch activity representation exhibits short-duration errors that could be easily addressed in a post-processing stage involving a temporal model, but also longer-duration errors, such as pitches below or above the restricted pitch range of the excerpt, that would be less easily handled. The pitch activity representation estimated from the fixed spectra involves even more errors. Although the learned basis spectra are clearly pitched, their spectral envelopes do not match those of the piano spectra in the test excerpt. Several pitches at integer fundamental frequency ratios are then combined to represent a single note.

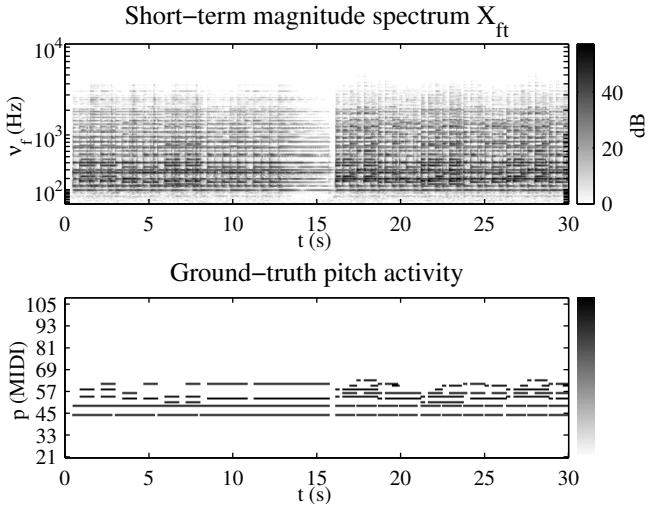


Figure 1: Magnitude spectrum and ground-truth pitch activity for the first 30 s of Borodin's *Little Suite - Serenade* for piano.

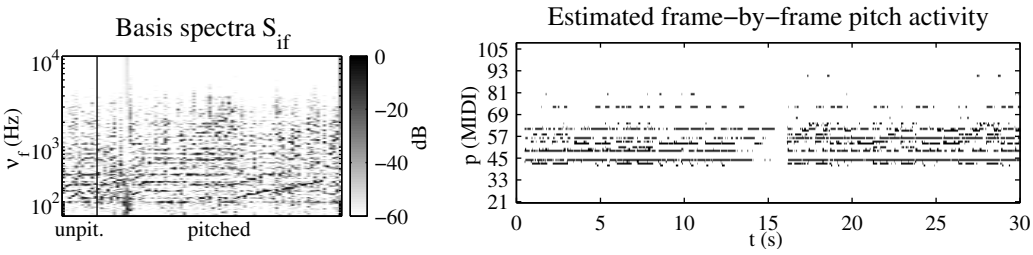


Figure 2: Basis spectra and pitch activity adaptively estimated for the piano excerpt in Fig. 1 via unconstrained NMF. The basis spectra were sorted by increasing estimated pitch.

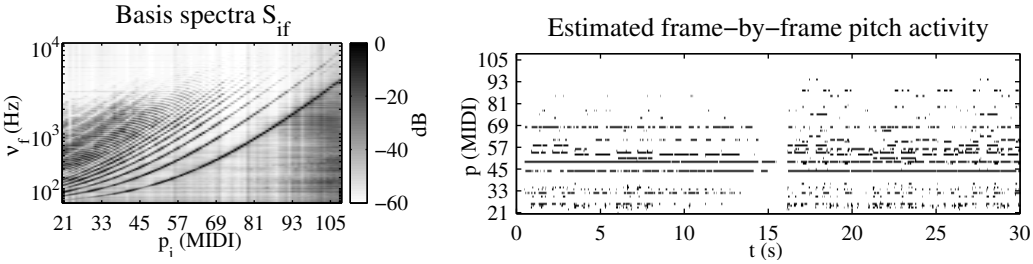


Figure 3: Basis spectra learned from the University of Iowa's piano samples and resulting pitch activity derived for the piano excerpt in Fig. 1 via NMF.

3 Adaptive harmonic decomposition

In order to avoid the above pitch estimation errors, it appears necessary to constrain each basis spectrum to represent a single note but to adapt its spectral envelope to the test data. We achieve these goals by adding constraints over the fine structure of the basis spectra within the model, but leaving some degrees of freedom over their spectral envelope.

3.1 General framework for spectral fine structure constraints

We associate each basis spectrum S_{if} with an integer pitch p and index by j, $1 \leq j \leq J_p$, the basis spectra having the same pitch but different spectral envelopes. The model spectrum (1) is then equivalently written as

$$X_{ft} = \sum_{p=p_{\text{low}}}^{p_{\text{high}}} \sum_{j=1}^{J_p} A_{pjt} S_{pjf}. \tag{9}$$

In order to ensure that each spectrum S_{pif} actually models the expected pitch p, we constrain it as

$$S_{pjf} = \sum_{k=1}^{K_p} E_{pjk} N_{pkf} \tag{10}$$

where N_{pkf}, $1 \leq k \leq K_p$, are fixed narrowband spectra enforcing the spectral fine structure associated with that pitch and the coefficients E_{pjk} parametrize the spectral envelope. The estimation of the model parameters now consists of inferring the spectral envelope and the time-varying amplitude of each basis spectrum from the test data, given its prior fine structure. Due to the linearity of constraint (10), the estimation of each of these two quantities can be recast into the standard NMF framework. The β-divergence can be minimized using the following multiplicative updates rules

$$A_{pjt} \leftarrow A_{pjt} \frac{\sum_{f=1}^{F} S_{pjf} Y_{ft}^{\beta-2} X_{ft}}{\sum_{f=1}^{F} S_{pjf} Y_{ft}^{\beta-1}} \tag{11}$$

$$E_{pjk} \leftarrow E_{pjk} \frac{\sum_{f=1}^{F} \sum_{t=1}^{T} A_{pjt} N_{pkf} Y_{ft}^{\beta-2} X_{ft}}{\sum_{f=1}^{F} \sum_{t=1}^{T} A_{pjt} N_{pkf} Y_{ft}^{\beta-1}} \tag{12}$$

whose convergence can be proved under the same conditions as above. In the following, we initialize the parameters prior to application of these rules by setting A_{pjt} to 1 and choosing E_{pjk} so that the basis spectra have a constant initial slope of $-6 \times j$ dB/octave over the whole frequency range regardless of their pitch.

3.2 Harmonicity and spectral smoothness constraints

The constraint (10) can represent a range of spectral fine structures associated with different instrument classes, including *e.g.* harmonic partials for woodwinds, slightly inharmonic partials for plucked strings or very inharmonic partials for bells. Given the frequencies of the partials, each fine structure spectrum N_{pkf} can be defined as a weighted sum of the spectra of individual partials

$$N_{pkf} = \sum_{m=1}^{M_p} W_{pkm} P_{pmf} \tag{13}$$

where P_{pmf} is the magnitude spectrum of the m-th overtone partial, M_p is the number of partials and the weights W_{pkm} parametrize the spectral shape of band k.

The spectrum of each partial can be analytically derived from the frequency responses of the bandpass filters associated with the frequency bins of the time-frequency transform. For the filterbank in Section 2.1, we get

$$P_{pmf} = \left| \mathrm{sinc}[L_f(\nu_f^{\mathrm{Hz}} - \nu_{pm}^{\mathrm{Hz}})] + \frac{1}{2}\mathrm{sinc}[L_f(\nu_f^{\mathrm{Hz}} - \nu_{pm}^{\mathrm{Hz}}) + 1] + \frac{1}{2}\mathrm{sinc}[L_f(\nu_f^{\mathrm{Hz}} - \nu_{pm}^{\mathrm{Hz}}) - 1] \right| \tag{14}$$

where ν_{pm}^{Hz} is the frequency of the m-th partial in Hz, sinc is the sine cardinal function and L_f is the length in seconds of the filter associated with bin f. We previously showed that the modeling of inharmonicity or variable tuning in this context does not significantly affect multiple pitch transcription performance on piano data compared to a harmonic model with fixed tuning [26]. Therefore we assume that the frequencies of the partials follow the exact harmonic model

$$\nu_{pm}^{\mathrm{Hz}} = m\, \nu_{p0}^{\mathrm{Hz}} \tag{15}$$

where the fundamental ν_{p0}^{Hz} corresponding to pitch p is defined in (5). All harmonics may be observed, hence the number of partials is set to $M_p = \lfloor \nu_F^{\mathrm{Hz}} / \nu_{p0}^{\mathrm{Hz}} \rfloor$ where $\lfloor . \rfloor$ denotes the floor function and ν_F^{Hz} the frequency of the topmost frequency bin.

The choice of the weights W_{pkm} in (13) affects pitch estimation performance. When each fine structure spectrum N_{pkf} represents a single partial, the basis spectra S_{pjf} may encode multiples of the expected fundamental frequency, resulting in substitution errors. When it contains too many partials, the basis spectra may not adapt well to the spectral envelope of the instruments, leading to insertion or deletion errors. In order to avoid such errors, each fine structure spectrum should span a narrow frequency band containing a few partials. The relative amplitudes of these partials may be chosen under the additional constraint of spectral smoothness, exploited by some other pitch estimation algorithms [7], enforcing similar amplitudes for adjacent partials. Practical

implementations of this constraint typically rely either on the properties of auditory pitch perception or those of musical instrument sounds.

We investigate a range of implementations by exploring different choices for the center frequencies, the bandwidths and the shapes of the fine structure spectra. The weights W_{pkm} are defined as

$$W_{pkm} = w\left(\frac{\nu_{pm} - \nu_{p0} - (k-1)b}{2b}\right) \tag{16}$$

where w is a chosen window function, ν_{p0} and ν_{pm} denote the frequency of the fundamental and that of the m-th partial on a chosen frequency scale, b is the spacing between successive frequency bands and $2b$ their bandwidth on that scale. The shape of the frequency bands is governed by w and their center frequencies are uniformly spaced on the chosen frequency scale, starting from the fundamental. The choice of a larger bandwidth $2b$ than the minimum bandwidth b needed for full coverage increases the smoothness of the resulting basis spectra. Similarly to above, all frequency bands are assumed to be observed up to a maximum index $K_{\max}$ so that the number of frequency bands is set to $K_p = \min(\lfloor(\nu_F - \nu_{p0})/b\rfloor + 1, K_{\max})$ with ν_F the frequency of the topmost frequency bin expressed on the chosen scale. The maximum total bandwidth is then equal to $b_{\max} = K_{\max}\, b$.

In the following, we consider three particular frequency scales: the pitch-synchronous linear scale indicating the partial index

$$\nu^{\text{psyn}} = \frac{\nu^{\text{Hz}}}{\nu^{\text{Hz}}_{p0}}, \tag{17}$$

the logarithmic octave scale

$$\nu^{\text{oct}} = \log_2 \nu^{\text{Hz}}, \tag{18}$$

and the ERB scale

$$\nu^{\text{ERB}} = 9.26 \log(0.00437\nu^{\text{Hz}} + 1). \tag{19}$$

In parallel, we consider four symmetric window functions of unitary bandwidth: the rectangular window

$$w^{\text{rect}}(u) = \begin{cases} 1 & \text{if } -\frac{1}{2} \leq u \leq \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

the triangular window

$$w^{\text{triang}}(u) = \begin{cases} 1 - |u| & \text{if } -1 \leq u \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

the Hann window

$$w^{\text{hann}}(u) = \begin{cases} \frac{1}{2}(1 + \cos \pi u) & \text{if } -1 \leq u \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

and the "gammatone" window of order n [33]

$$w^{\text{gamma}}(u) = \frac{1}{(1 + k^2u^2)^n} \text{ with } k = \frac{\sqrt{\pi}\,\Gamma(n - 1/2)}{\Gamma(n)} \tag{23}$$

with $\Gamma(.)$ denoting the gamma function. By contrast with other windows, the latter has infinite support and allows control of the rolloff slope via its parameter n.

The ERB scale and the gammatone window are both perceptually motivated [33]. The spectral envelope coefficients E_{pjk} corresponding to these choices are hence closely related to the frequency-warped cepstral coefficients routinely used as timbre features for audio classification [34]. Example spectra corresponding to these choices are shown in Fig. 4. Although audiological measurements suggest that the shape of auditory bands is asymmetric on the ERB scale, we observed that the use of symmetric windows did not significantly affect pitch estimation performance. A similar model involving triangular windows with a spacing and a bandwidth of 2/3 octave was employed in [35] for the estimation of the amplitudes of overlapping partials given estimated pitches.

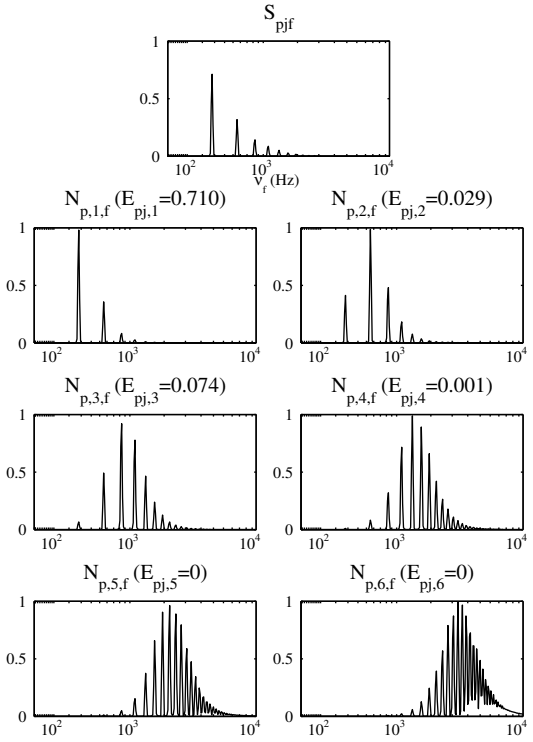


Figure 4: Basis spectrum S_{pjf} estimated for the piano excerpt in Fig. 1 given fixed harmonic fine structure spectra N_{pkf} ($p = 60$, gammatone windows of order $n = 4$, $b = 11/3$ ERB, $K_{\max} = 6$).

3.3 Example results

Fig. 5 depicts the pitch estimates obtained via NMF under harmonicity and spectral smoothness constraints on the piano excerpt in Fig. 1 given a pitch activity detection threshold $A_{\min}$ of 25 dB. Comparison with Figs. 2 and 3 indicates that these estimates are more accurate. In particular, the number of short-duration errors is decreased and the estimated pitches lie mostly within the the true pitch range of the excerpt. Some basis spectra, *e.g.* around $p = 80$, are inaccurately estimated due to the lack of observed data corresponding to these pitches. However this does not reflect in the estimated pitches.

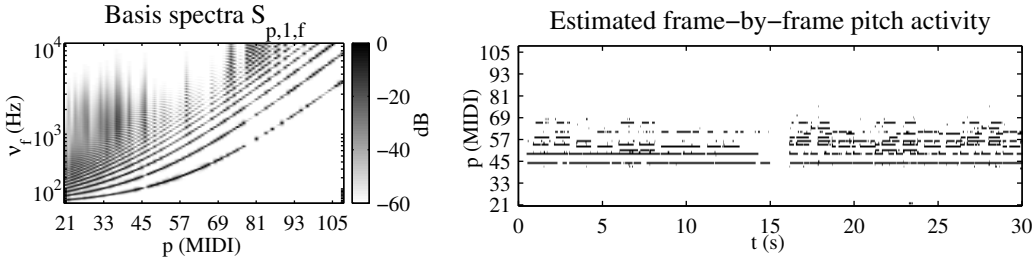


Figure 5: Basis spectra and pitch activity estimated for the signal in Fig. 1 via NMF under harmonicity and spectral smoothness constraints (implemented with gammatone windows of order $n = 4$, $b = 11/3$ ERB, $K_{\max} = 6$).

3.4 Learning the fine structure

An alternative approach to the definition of the fine structure spectra N_{pkf} not relying on harmonicity and spectral smoothness assumptions is to train them on annotated samples of several instruments sharing similar spectral fine structures. In order to ensure that the learned spectra exhibit a narrow bandwidth, their frequency

support can be constrained similarly to above via

$$N_{pkf} = 0 \text{ if } |\nu_f - \nu_{p0} - (k-1)b| > 2b \tag{24}$$

where ν_f and ν_{p0} are the frequency of bin f and the fundamental frequency measured over one of the frequency scales in (17), (18), (19), b is the spacing between successive frequency bands and $2b$ their bandwidth on that scale. The training objective can again be recast into the standard NMF framework, leading to the multiplicative update rule

$$N_{pkf} \leftarrow N_{pkf} \frac{\sum_{j=1}^{J_p} \sum_{t=1}^{T} A_{pjt} E_{pjk} Y_{ft}^{\beta-2} X_{ft}}{\sum_{j=1}^{J_p} \sum_{t=1}^{T} A_{pjt} E_{pjk} Y_{ft}^{\beta-1}} \tag{25}$$

to be applied alternatingly with (11) and (12). By property of multiplicative updates, the constraint (24) remains true at each iteration provided it is initially satisfied.

4 Evaluation

4.1 Algorithms and evaluation metrics

We evaluated the algorithms in Sections 2 and 3 on two distinct datasets: a subset of the MAPS piano database [31] and the woodwind training dataset for the Multiple Fundamental Frequency Estimation task of the Third Music Information Retrieval Evaluation eXchange[1] (MIREX 2007). Algorithms based on fixed spectra were trained on isolated piano sounds from the MIS database [32] and the RWC Musical Instrument Sound Database [36], which cover the full pitch range at three loudness levels of one and three pianos, respectively.

Two additional NMF algorithms were tested for comparison: NMF under harmonicity and source-filter constraints [22] and NMF under a single harmonicity constraint identical to that in [23] except for the improved modeling of the partial spectra in (14). The distortion measure used in the original algorithms was replaced by the more general β-divergence and optimized via multiplicative updates initialized in the same way as other NMF algorithms, *i.e.* with a -6 dB/octave slope for the harmonic spectra and a flat slope for the filter. The harmonic sum algorithm in [7], provided by its author, and the piano-specific SONIC algorithm in [8][2], which estimates the onset and duration of each note, were also evaluated. In order to allow fair comparison despite the use of a different front-end, the frame size of the harmonic sum algorithm was set to 46 ms, which is close to the effective time resolution of the ERB filterbank at the fundamental frequency corresponding to the average observed pitch.

The accuracy of the amplitude-based pitch salience measures produced by all algorithms except SONIC was assessed by interpolating them over 10 ms time frames and deriving pitch estimates as explained in Section 2.4. Frame-by-frame pitch estimates were also derived for SONIC from the onset and duration of each estimated note. Performance was quantified in terms of recall $\mathcal{R}$, precision $\mathcal{P}$ and F-measure $\mathcal{F}$ [37] and averaged over all items within each dataset.

4.2 Results on piano data

The first dataset consists of the initial 30 s of 50 piano pieces from the MAPS database, recorded from a Disklavier acoustic piano using either close or ambiance microphones, and having a polyphony level of 3.9 on average and 9 at most. Due to the lack of sufficient annotated data from different pianos, the optimal parameter values for each algorithm were not learned a priori. Instead, we considered a range of values and analyzed the impact on performance of each parameter, other parameters being fixed to their optimal values. Although the optimal a posteriori performance figures are presumably larger than with prior parameter settings, we believe that this allows fair comparison of algorithms in terms of relative performance, as well as deeper understanding of the sensitivity to each parameter. For all NMF algorithms, the number of basis spectra was chosen among multiples of 88, the distortion measure parameter β was varied between 0 and 2 in steps of 0.1 and the detection threshold $A_{\min}$ between 15 and 40 dB in steps of 1 dB. For the proposed NMF algorithm, preliminary experiments showed that, although the effect on performance of the maximum number of frequency bands $K_{\max}$ and their bandwidth b are related, that of $K_{\max}$ and the maximum total bandwidth $B_{\max}$ are roughly independent. The latter was varied in steps of 1 partial, 1/3 octave or 2 ERB, depending on the chosen frequency scale, and b was derived as $b = B_{\max}/K_{\max}$.

[1] http://www.music-ir.org/mirex2007/
[2] http://lgm.fri.uni-lj.si/sonic.html

Table 1: Average pitch estimation performance over piano data using optimal parameter values for each algorithm.

Algorithm	$\mathcal{P}$ (%)	$\mathcal{R}$ (%)	$\mathcal{F}$ (%)
No training			
Unconstrained NMF	58.9	60.0	57.8
NMF under harmonicity constraint	63.2	60.9	60.5
NMF under harmonicity and source-filter constraints [22]	60.1	59.1	57.5
NMF under harmonicity and spectral smoothness constraints	71.6	65.5	67.0
Harmonic sum [7]	65.7	57.4	60.2
Training on piano data			
NMF with basis spectra trained on MIS	61.2	62.1	59.6
NMF with basis spectra trained on MIS & RWC	68.6	66.7	66.0
NMF with fine structure spectra trained on MIS & RWC	67.2	64.9	64.2
Training on piano data and post-processing			
SONIC [8]	74.5	57.6	63.6

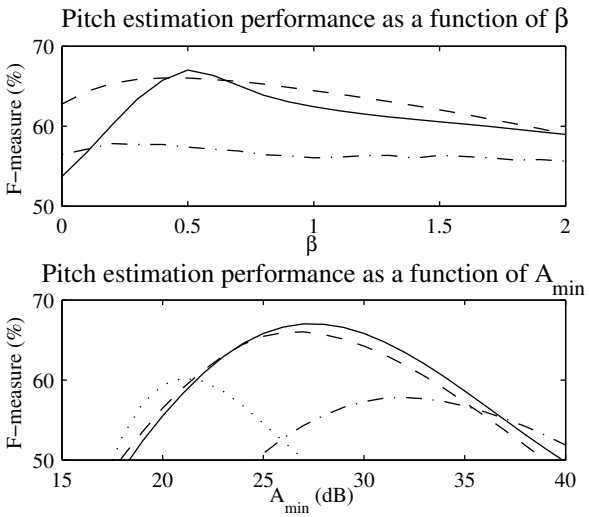


Figure 6: Variation of the average pitch estimation performance over piano data as a function of the divergence parameter β and the detection threshold A_{min}. Plain: NMF under harmonicity and spectral smoothness constraints, dashed: NMF with basis spectra trained on MIS & RWC, dash-dotted: unconstrained NMF, dotted: harmonic sum [7].

The results with the optimal parameter values are given in Table 1. The proposed algorithm with fixed fine structure spectra yields an average F-measure of 67%. This is comparable to the performance of NMF with fixed spectra learned on both MIS and RWC, but about 9% better than unconstrained NMF, 6% better than NMF under harmonicity constraint alone, 10% better than NMF under harmonicity and source-filter constraints and 7% better than the harmonic sum algorithm. This confirms that harmonicity is an appropriate but insufficient constraint in the context of pitch estimation and suggests that spectral smoothness is more useful than source-filter modeling as an additional constraint. Fine structure spectra learned on piano data did not further improve performance compared to fixed fine structure spectra.

For all NMF algorithms, the F-measure was maximum with $I = 88$ basis spectra and decreased by 1 to 5% with $I = 176$ and 2 to 7% with $I = 264$. Performance variation as a function of β and A_{min} is depicted in Fig. 6. As explained in [19], a small value of β appears preferable for unconstrained NMF in order to infer wideband spectral structures despite the wide differences in dynamics between low and high frequencies. For

Table 2: Variation of the average pitch estimation performance over piano data of NMF under harmonicity and spectral smoothness constraints for different frequency scales.

Frequency scale	Optimal parameters	$\mathcal{F}$ (%)
Pitch-synchronous	Gammatone $n = 2$ $K_{\max} = 6$ $B_{\max} = 6$ partials	66.1
Octave	Gammatone $n = 4$ $K_{\max} = 5$ $B_{\max} = 13/3$ octaves	66.5
ERB	Gammatone $n = 4$ $K_{\max} = 6$ $B_{\max} = 22$ ERB	67.0

Table 3: Variation of the average pitch estimation performance over piano data of NMF under harmonicity and spectral smoothness constraints for different band shapes.

Window function w	$\mathcal{F}$ (%)
Rectangular	60.7
Triangular	64.4
Hann	63.8
Gammatone $n = 2$	66.0
Gammatone $n = 4$	67.0
Gammatone $n = 6$	66.5

other algorithms, the optimal β is equal to 0.5. The resulting distortion measure scales similarly to perceptual loudness for audible sounds and was also shown to be optimal in the context of audio source separation in [28]. Doubling or halving β decreases the F-measure by 0 to 5%. Unconstrained NMF also exhibits a distinct behavior from other NMF algorithms when considering the choice of $A_{\min}$, with an optimal value of 32 dB instead of a more conservative 27 dB. A deviation of 3 dB from the optimal $A_{\min}$ decreases the F-measure by 1 to 2%. The harmonic sum algorithm is more sensitive to the choice of $A_{\min}$, with a decrease up to 7% for the same deviation.

The best results for the proposed algorithm were obtained when building fine structure spectra from gammatone windows of order $n = 4$ spaced on the ERB scale, with a maximum number of $K_{\max} = 6$ frequency bands and a maximum total bandwidth $B_{\max} = 22$ ERB. The effect of these parameters is analyzed in Tables 2 and 3 and in Fig. 7. The frequency scale has little influence, provided other parameters are adapted to the chosen scale. The bandwidth of each spectrum also has little influence, since any value of $K_{\max}$ between 4 and 11 or any value of $B_{\max}$ larger than 18 ERB results in an average F-measure within 2% of the optimum. Small values of $K_{\max}$ and $B_{\max}$ should be avoided, since they result in insufficient adaptation capabilities or incomplete coverage of the frequency axis, respectively. Finally, gammatone windows perform about 3% better than smooth windows with finite support, but the window order is not critical. Only rectangular windows should be avoided. Overall, this suggests that, even if it is not optimally implemented, the spectral smoothness constraint still improves performance compared to the harmonicity constraint alone, provided the window w is smooth and $K_{\max}$ and $B_{\max}$ are large enough.

4.3 Results on woodwind data

After determining the optimal parameter values in Section 4.2, we applied some algorithms to a second dataset. From the recordings of individual instrument parts of a woodwind quintet by Beethoven made available at MIREX 2007, we generated four test excerpts with two to five instruments by successively summing together the initial 30 s of the parts of flute, clarinet, bassoon, horn and oboe. Pitch estimation results are listed in Table 4. NMF under harmonicity and spectral smoothness constraints performed best for most polyphonies, while NMF under harmonicity constraint alone sometimes performed worse than unconstrained NMF. Despite the fact that some pitches were played by up to three instruments, performance did not improve when employing more than one basis spectrum per pitch. Further experiments suggest that this is due both to the use of a constant number of basis spectra per pitch and to the difficulty of initializing these spectra so that each converges to a particular instrument.

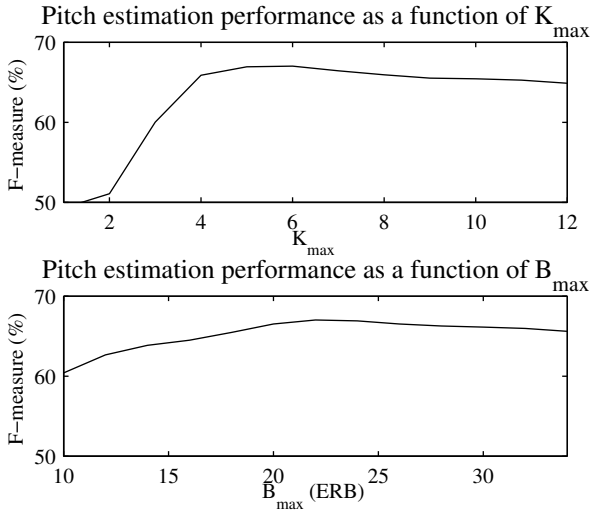


Figure 7: Variation of the average pitch estimation performance over piano data of NMF under harmonicity and spectral smoothness constraints as a function of the maximum number of frequency bands $K_{\max}$ and the maximum total bandwidth $B_{\max}$.

Table 4: F-measure (%) for pitch estimation over woodwind data.

Algorithm	Polyphony			
	2	3	4	5
Unconstrained NMF	79.9	56.3	62.1	61.9
NMF under harmonicity constraint	78.7	57.3	57.1	56.5
NMF under harmonicity and spectral smoothness constraints	76.5	64.7	67.5	62.5
Harmonic sum [7]	73.4	59.1	63.5	59.9

5 Conclusion

We proposed an adaptive spectral decomposition model for music signals based on harmonicity and spectral smoothness constraints. This model ensures that the estimated basis spectra have a known fine structure, while their spectral envelope is adapted to the observed data. Multiple pitch estimation experiments conducted on piano and woodwind data indicate that, independently of any temporal prior, the resulting constrained NMF algorithm is potentially competitive with NMF based on fixed instrument-specific spectra and superior to unconstrained NMF or NMF under harmonicity constraint alone. As a side result, we provided a benchmark of classical NMF algorithms in the context of multiple pitch estimation and showed that the optimal value of the β-divergence parameter is often different from the integer values commonly used in the literature.

In the future, we plan to exploit the estimated amplitude-based pitch salience measure for music-to-score transcription via a probabilistic model involving additional temporal priors. Given their relationship to frequency-warped cepstral coefficients, the estimated spectral envelope coefficients could then be used to cluster the notes into instrument parts. We also aim to extend our model to represent percussive as well as pitched instruments and to improve its performance over mixtures of several instruments by using an adaptive number of basis spectra per pitch, based on recent findings regarding the estimation of the number of basis spectra [38] and their initialization [39].

Acknowledgments

We would like to thank Anssi Klapuri for sharing the code of the harmonic sum algorithm, Valentin Emiya for providing information about the MAPS database and MIDI handling in Matlab and Mert Bay for generating the woodwind dataset.

References

[1] A.P. Klapuri and M. Davy, *Signal processing methods for music transcription*, Springer, New York, NY, 2006.

[2] M.P. Ryynänen and A.P. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 319 – 322.

[3] M.P. Ryynänen and A.P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.

[4] J. Eggink and G.J. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2003, pp. 125–131.

[5] M.R. Every and J.E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1845–1856, 2006.

[6] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[7] A.P. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2006, pp. 216–221.

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

[9] J.P. Bello, L. Daudet, and M.B.Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2242–2251, 2006.

[10] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of western tonal music," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498–2517, 2006.

[11] G.E. Poliner and D.P.W. Ellis, "A discriminative model for polyphonic piano transcription," *Eurasip Journal of Advances in Signal Processing*, vol. 2007, 2007, Article ID 48317.

[12] D. FitzGerald, M. Cranitch, and E. Coyle, "Generalised prior subspace analysis for polyphonic pitch transcription," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2005.

[13] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[14] A. Cont, "Realtime multiple pitch observation using sparse non-negative constraints," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2006, pp. 206–212.

[15] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[16] S.A. Abdallah and M.D. Plumbley, "Unsupervised analysis of polyphonic music using sparse coding," *IEEE Trans. on Neural Networks*, vol. 17, no. 1, pp. 179–196, 2006.

[17] 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. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 1, pp. 65–68.

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

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, 2009, In press.

[20] M. Kim and S. Choi, "Monaural music source separation: nonnegativity, sparseness and shift-invariance," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2006, pp. 617–624.

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

[22] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008, Article ID 872425.

[23] S.A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2007, pp. 381–386.

[24] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 169–172.

[25] E. Vincent, N. Bertin, and R. Badeau, "Two nonnegative matrix factorization methods for polyphonic pitch transcription," in *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

[26] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 109–112.

[27] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models, 2nd Edition*, Springer, Heidelberg, 1999.

[28] P.D. O'Grady, *Sparse separation of under-determined speech mixtures*, Ph.D. thesis, National University of Ireland Maynooth, 2007.

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

[30] P.M. Brossier, *Automatic annotation of musical audio for interactive applications*, Ph.D. thesis, Centre for Digital Music, Queen Mary University of London, UK, 2007.

[31] V. Emiya, *Transcription automatique de la musique de piano*, Ph.D. thesis, TELECOM ParisTech, France, 2008.

[32] The University of Iowa Electronic Music Studios, "Musical instrument samples," http://theremin.music.uiowa.edu/MIS.html.

[33] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psycho-acoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1805–1808.

[34] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[35] T. Virtanen and A.P. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1757–1760.

[36] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2003, pp. 229–230.

[37] C.J. van Rijsbergen, *Information retrieval, 2nd Edition*, Butterworths, London, UK, 1979.

[38] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," Tech. Rep. CUED/F-INFENG/TR.609, University of Cambridge, UK, 2008.

[39] Z. Zheng, J. Yang, and Y. Zhu, "Initialization enhancer for non-negative matrix factorization," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 1, pp. 101–110, 2007.