Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle

Valentin Emiya, *Member, IEEE,* Roland Badeau, *Member, IEEE,* Bertrand David, *Member, IEEE*

***Abstract*—A new method for the estimation of multiple concurrent pitches in piano recordings is presented. It addresses the issue of overlapping overtones by modeling the spectral envelope of the overtones of each note with a smooth autoregressive model. For the background noise, a moving-average model is used and the combination of both tends to eliminate harmonic and sub-harmonic erroneous pitch estimations. This leads to a complete generative spectral model for simultaneous piano notes, which also explicitly includes the typical deviation from exact harmonicity in a piano overtone series. The pitch set which maximizes an approximate likelihood is selected from among a restricted number of possible pitch combinations as the one. Tests have been conducted on a large homemade database called MAPS, composed of piano recordings from a real upright piano and from high-quality samples.**

***Index Terms*—Acoustic signal analysis, audio processing, multipitch estimation, piano, transcription, spectral smoothness.**

I. Introduction

The issue of monopitch estimation has been addressed for decades by different approaches such as retrieving a periodic pattern in a waveform [1; 2] or matching a regularly spaced pattern to an observed spectrum [3–5], or even by combining both spectral and temporal cues [6; 7]. Conversely the multipitch estimation (MPE) problem has become a rather active research area in the last decade [8–11] and is mostly handled by processing spectral or time-frequency representations. This MPE task has also become a central tool in musical scene analysis [11], particularly when the targeted application is the automatic transcription of music (ATM) [12–16]. While recent works consider pitch and time dimensions jointly to perform this task, MPE on single frames has been historically used as a prior processing to pitch tracking over time and musical note detection. Indeed, in a signal processing perspective, a period or a harmonic series can be estimated from a short signal snapshot. In terms of perception, only a few cycles are needed to identify a pitched note [17].

In this polyphonic context, two issues have proved difficult and interesting for computational MPE: the overlap between the overtones of different notes and the unknown number of such notes occurring simultaneously. As the superposition of two sounds in octave relationship leads to a spectral ambiguity and thus to an ill-posed problem, some knowledge is often used for the further spectral modeling of the underlying sources. For instance, when iterative spectral subtraction is employed, several authors have adopted a smoothness assumption for the spectral envelope of the notes [8–10].

Several preceding works [12–14; 16] specifically address the case of the piano. One of the interests in studying a single instrument is to incorporate in the algorithms some results established from physical considerations [18; 19], in the hope that a more specific model will lead to a better performance, with the obvious drawback of narrowing the scope of the application. Indeed, while a large number of musical pieces are composed for piano solo, the performance of ATM systems is relatively poor for this instrument in comparison to others [6].

Two deviations from a simple harmonic spectral model are specific to free vibrating string instruments and are particularly salient for the piano: a small departure from exact harmonicity (the overtone series is slightly stretched) and the beatings between very close frequency components (for a single string two different vibrating polarizations occur and the different strings in a doublet or triplet are slightly detuned on purpose [20]).

In this work, we propose a new spectral model in which the inharmonic distribution is taken into account and adjusted for each possible note. The spectral envelope of the overtones is modeled by a smooth autoregressive (AR) model, the smoothness resulting from a low model order. A smooth spectral model is similarly introduced for the residual noise by using a low-order moving-average (MA) process, which is particular efficient against residual sinusoids. The proposed method follows a preliminary work [21] in which the AR/MA approach was already used and MPE was addressed as an extension of a monopitch estimation approach. In addition to the opportunity to deal with higher polyphony levels, the current paper introduces a new signal model for simultaneous piano notes and a corresponding estimation scheme. The major advance from the previous study is thus to model and estimate the spectral overlap between note spectra. An iterative estimation framework is proposed which leads to a maximum likelihood estimation for each possible set of F_0s.

This paper is structured as follows. In section II, the overall MPE approach is described. The sound model is first detailed, followed by the principle of the algorithm: statistical background, adaptive choice of the search space and detection function. The estimation of the model parameters is then explained in section III. After specifying the implementation details, the algorithm is tested in section IV using a database of piano sounds and the results are analyzed and compared with those obtained with some state-of-the-art approaches.

V. Emiya is with the Metiss team at INRIA, Centre Inria Rennes - Bretagne Atlantique, Rennes, France, and, together with R. Badeau and B. David, with Institut Télécom; Télécom ParisTech; CNRS LTCI, Paris, France.

The research leading to this paper was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, *Décomposition en Éléments Sonores et Applications Musicales - DESAM* and by the European Commission under contract FP6-027026-K-SPACE. The authors would also like to thanks Peter Weyer-Brown from Telecom ParisTech and Nathalie Berthet for their useful comments to improve the English usage.

Conclusions are finally drawn in section V.

Note that in the following, * denotes the complex conjugate, $^{\mathrm{T}}$ the transpose operator, † the conjugate transpose, $\lfloor . \rfloor$ the floor function and $|.|$ the number of elements in a set, the absolute value of a real number or the modulus of a complex number. In addition, the term *polyphony* will be applied to any possible mixture of notes, including silence (*polyphony 0*) and single notes (*polyphony 1*).

II. Multipitch estimation algorithm

A. Generative sound model

At the frame level, a mixture of piano sounds is modeled as a sum of sinusoids and background noise. The amplitude of each sinusoid is considered as a random variable. The spectral envelope for the overtone series of a note is introduced as second order statistical properties of this random variable, making it possible to adjust its smoothness. The noise is considered as a moving-average (MA) process.

More precisely, a mixture of $P \in \mathbb{N}$ simultaneous piano notes, observed in a discrete-time N-length frame, is modeled as a sum $x(t) \triangleq \sum_{p=1}^{P} x_p(t) + x_b(t)$ of the note signals x_p and of noise x_b. The sinusoidal model for x_p is

$$x_p(t) \triangleq \sum_{h=1}^{H_p} \left(\alpha_{hp} e^{2i\pi f_{hp} t} + \alpha_{hp}^* e^{-2i\pi f_{hp} t}\right) \quad (1)$$

where α_{hp} are the complex amplitudes and f_{hp} the frequencies of the H_p overtones (here, H_p is set to the maximum number of overtones below the Nyquist frequency). Note p is parameterized by $\mathcal{C}_p = (f_{0p}, \beta_p)$, f_{0p} being the *fundamental frequency* (F_0) and $\beta_p(f_{0p})$ being the so-called *inharmonicity coefficient* of the piano note [20], such that

$$f_{hp} \triangleq h f_{0p} \sqrt{1 + \beta_p(f_{0p})\, h^2} \quad (2)$$

We introduce a *spectral envelope* for note p as an autoregressive (AR) model of order Q_p, parameterized by $\theta_p \triangleq \left(\sigma_p^2, A_p(z)\right)$, where σ_p^2 is the power and $\frac{1}{A_p(z)}$ is the transfer function of the related AR filter. Order Q_p should be low and proportional to H_p, in order to obtain a smooth envelope and to avoid overfitting issues for high pitches. The *amplitude* α_{hp} of overtone h of note p is the outcome of a zero-mean complex Gaussian random variable[1], with variance equal to the spectral envelope power density at the frequency f_{hp} of the overtone:

$$\alpha_{hp} \sim \mathcal{N}\left(0, \frac{\sigma_p^2}{\left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2}\right) \quad (3)$$

In order to model a smooth spectral envelope without residual peaks, the *noise* x_b is modeled as an MA process with a low order Q_b, parameterized by $\theta_b \triangleq \left(\sigma_b^2, B(z)\right)$, σ_b^2 being the power and $B(z)$ the finite impulse response (FIR) filter of the process.

Furthermore, high note powers and low noise powers are favored by choosing an inverse gamma prior for σ_p^2: $\sigma_p^2 \sim \mathcal{IG}\left(k_{\sigma_p^2}, E_{\sigma_p^2}\right)$ and a gamma prior for σ_b^2: $\sigma_b^2 \sim \Gamma\left(k_{\sigma_b^2}, E_{\sigma_b^2}\right)$. A non-informative prior[2] is assumed for $A_p(z)$ and $B(z)$.

[1]Using these centered Gaussian variables, x_p is thus a harmonic process, in the statistical sense.

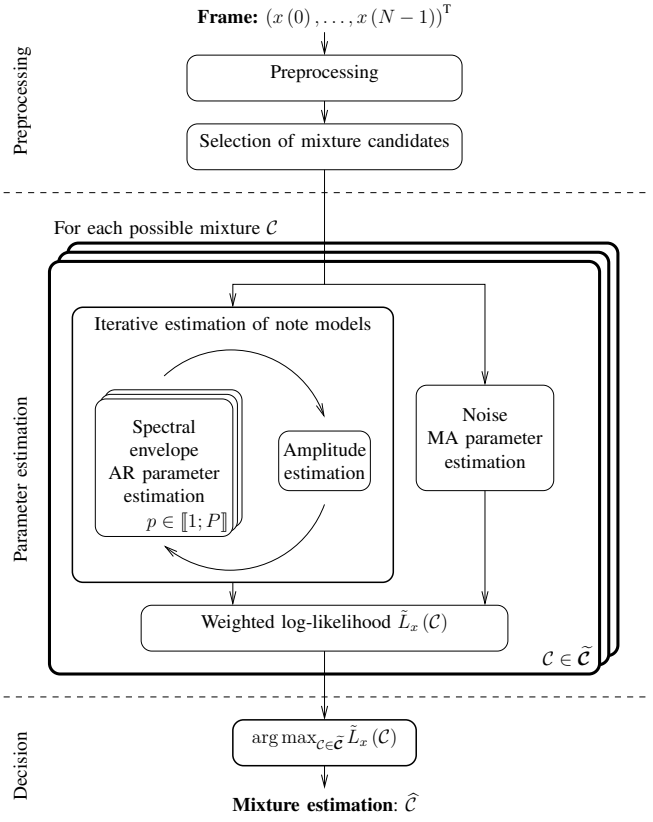


Fig. 1. Multipitch estimation block diagram.

With a user-defined weighting window w, we can equivalently consider the Discrete Time Fourier Transforms (DTFT) $X_p(f)$ of $x_p(t)\,w(t)$, $X_b(f)$ of $x_b(t)\,w(t)$ and $X(f) \triangleq \sum_{p=1}^{P} X_p(f) + X_b(f)$ of $x(t)\,w(t)$. W being the DTFT of w, we thus have

$$X_p(f) = \sum_{h=1}^{H_p} \left(\alpha_{hp} W(f - f_{hp}) + \alpha_{hp}^* W^*(f + f_{hp})\right) \quad (4)$$

In addition, the following synthetic notations will be used: a mixture of notes is denoted by $\mathcal{C} \triangleq (\mathcal{C}_1, \ldots, \mathcal{C}_P)$; the set of parameters for spectral envelope models by $\theta \triangleq (\theta_1, \ldots, \theta_P)$; the set of amplitudes of overtones of note p by $\alpha_p \triangleq \left(\alpha_{1p}, \ldots, \alpha_{H_p p}\right)$; and the set of amplitudes of overtones of all notes by $\alpha \triangleq (\alpha_1, \ldots, \alpha_P)$.

B. Principle of the algorithm

The algorithm is illustrated in Fig. 1. The main principles are described in the current section, except for the model parameter estimation, which is detailed in Section III.

[2]*e.g.* an improper, constant prior density, which will not affect the detection function.

1) Statistical background: Let us consider the observed frame x and the set $\boldsymbol{\mathcal{C}}$ of all the possible mixtures of piano notes. Ideally, the multipitch detection function can be expressed as the maximum *a posteriori* (MAP) estimator $\hat{\mathcal{C}}$, which is equivalent to the ML estimator if no *a priori* information on mixtures is specified:

$$\begin{aligned}\hat{\mathcal{C}} &= \arg\max_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} p(\mathcal{C}|x) = \arg\max_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} \frac{p(x|\mathcal{C})\,p(\mathcal{C})}{p(x)} \\ &= \arg\max_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} p(x|\mathcal{C}) \end{aligned} \tag{5}$$

2) Search space: The number of combinations among Q possible notes being $\binom{Q}{P}$ for polyphony P, the size of the search set $\boldsymbol{\mathcal{C}}$ is $\sum_{P=0}^{Q}\binom{Q}{P} = 2^Q$, *i.e.* around $3{\cdot}10^{26}$ for a typical piano with $Q = 88$ keys. Even when the polyphony is limited to $P_{\max} = 6$ and the number of possible notes to $Q = 60$ (*e.g.* by ignoring the lowest and highest-pitched keys), the number of combinations reaches $\sum_{P=0}^{P_{\max}}\binom{Q}{P} \approx 56{\cdot}10^6$, which remains a too huge search set for a realistic implementation. In order to reduce the size of this set, we propose to select a given number N_c of note candidates. We use the normalized product spectrum function defined in dB as

$$\Pi_X(f_0,\beta) \triangleq \frac{1}{H(f_0,\beta)^{\nu}} 10\log \prod_{h=1}^{H(f_0,\beta)} |X(f_h)|^2 \tag{6}$$

where X is the observed spectrum, $H(f_0,\beta)$ is the number of overtones for the note with fundamental frequency f_0 and inharmonicity β, $f_h \triangleq hf_0\sqrt{1+\beta h^2}$, and ν is a parameter adjusted to balance the values of the function between bass and treble notes. As the true notes generate local peaks in Π_X, selecting an oversized set of N_c greatest peaks is an efficient way for adaptively reducing the number of note candidates. In addition, for candidate $n_c \in [\![1; N_c]\!]$, accurate values of its fundamental frequency f_{0n_c} and inharmonicity β_{n_c} are found by a local two-dimensional maximization[3] of $\Pi_X(f_{0n_c},\beta_{n_c})$. They are used in the subsequent estimation stages to locate the frequencies of the overtones. In the current implementation, the number of note candidates is set to $N_c = 9$. This choice results from the balance between increasing the number of candidates and limiting the computational time. The size of the set $\widetilde{\boldsymbol{\mathcal{C}}}$ of possible mixtures thus equals $\sum_{P=0}^{P_{\max}}\binom{N_c}{P} = 466$, for $P_{\max} = 6$.

3) Multipitch detection function: The multipitch detection function aims at finding the correct mixture among all the possible candidates. For each candidate $\mathcal{C} \in \widetilde{\boldsymbol{\mathcal{C}}}$, the parameters $\widehat{\alpha}$, $\widehat{\theta}$ and $\widehat{\theta}_b$ of the related model are estimated as explained in section III. The detection function is then computed, and the multipitch estimation is defined as the mixture with the greatest value.

Since equation (5) is intractable, we define an alternate detection function which involves the following terms:

- $L_p(\theta_p) \triangleq \ln p(\alpha_p|\theta_p,\mathcal{C}_p)$ is the log-likelihood of the amplitudes α_p of overtones of note p, related to the spectral envelope models θ_p;
- $L_b(\theta_b) \triangleq \ln p(x|\alpha,\theta_b,\mathcal{C})$ is equal[4] to the log-likelihood $\ln p_{x_b}$ related to the noise model θ_b;
- the priors $\ln p(\theta_p)$ on the spectral envelope of the note p and $\ln p(\theta_b)$ on noise parameters.

We empirically define the detection function as a weighted sum of these log-densities, computed with the estimated parameters (see Appendix A for a discussion):

$$\begin{aligned}\tilde{L}_x(\mathcal{C}) \triangleq\ & w_1 \sum_{p=1}^{P} \tilde{L}_p\left(\widehat{\theta_p}\right)/P + w_2 \tilde{L}_b\left(\widehat{\theta_b}\right) \\ & + w_3 \sum_{p=1}^{P} \ln p\left(\widehat{\sigma_p^2}\right)/P + w_4 \ln p\left(\widehat{\sigma_b^2}\right) - \mu_{\text{pol}}\,P \end{aligned} \tag{7}$$

where

$$\begin{cases} \tilde{L}_p\left(\widehat{\theta_p}\right) \triangleq \frac{1}{H_p} L_p\left(\widehat{\sigma_p^2}, \widehat{A_p}\right) - \mu_{\text{env}}\,H_p \\ \tilde{L}_b\left(\widehat{\theta_b}\right) \triangleq \frac{1}{|\mathcal{F}_b|} L_b\left(\widehat{\sigma_b^2}, \widehat{B}\right) - \mu_b\,|\mathcal{F}_b| \\ w_1,\ldots,w_4,\mu_{\text{pol}},\mu_{\text{env}},\mu_b \text{ are user-defined} \\ \quad \text{coefficients} \\ |\mathcal{F}_b| \text{ is the number of noisy bins (see eq. (14)).} \end{cases}$$

III. Model parameter estimation

A. Iterative estimation of spectral envelope parameters and amplitudes of notes

We consider a possible mixture $\mathcal{C} = (\mathcal{C}_1,\ldots,\mathcal{C}_P)$. The estimation of the unknown spectral envelope parameters θ and of the unknown amplitudes of the overtones α is performed by iteratively estimating the former and the latter, as described in this section.

1) Spectral envelope parameter estimation: let us assume the amplitudes α are known in order to estimate the spectral envelope parameters θ in the maximum likelihood (ML) sense. Given that note models are independent, the likelihood of α is $p(\alpha|\theta,\mathcal{C}) = \prod_{p=1}^{P} p(\alpha_p|\theta_p,\mathcal{C}_p)$. The optimization w.r.t. θ thus consists in maximizing the log-likelihood $L_p\left(\sigma_p^2, A_p\right) \triangleq \ln p\left(\alpha_p|\sigma_p^2, A_p, \mathcal{C}_p\right)$ w.r.t. $\theta_p = \left(\sigma_p^2, A_p\right)$, independently for each note p. As proved in the Appendix B, the maximization w.r.t. σ_p^2 leads to the expression

$$L_p\left(\widehat{\sigma_p^2}, A_p\right) = c + \frac{H_p}{2}\ln\rho(A_p) \tag{8}$$

with

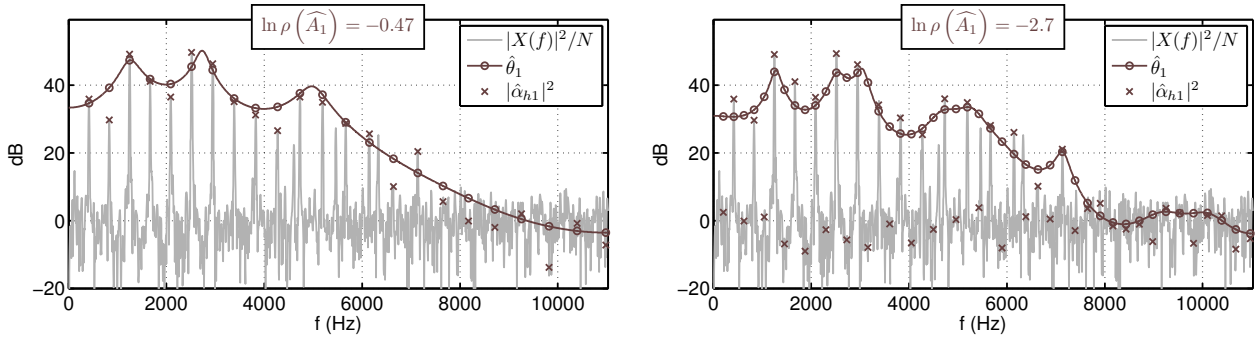
$$c \triangleq -\frac{H_p}{2}\ln(2\pi e) - \frac{1}{2}\sum_{h=1}^{H_p}\ln|\alpha_{hp}|^2 \tag{9}$$

$$\rho(A_p) \triangleq \frac{\left(\prod_{h=1}^{H_p}|\alpha_{hp}|^2\left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2\right)^{\frac{1}{H_p}}}{\frac{1}{H_p}\sum_{h=1}^{H_p}|\alpha_{hp}|^2\left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2} \tag{10}$$

$$\widehat{\sigma_p^2} = \frac{1}{H_p}\sum_{h=1}^{H_p}|\alpha_{hp}|^2\left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2 \tag{11}$$

[3]The `fminsearch` Matlab function was used for this optimization.

[4]by using the substitution $x \mapsto x - 2\text{Re}\left(\sum_{p=1}^{P}\sum_{h=1}^{H_p}\alpha_{hp}e^{2i\pi f_{hp}\mathbf{t}}\right)$, $\mathbf{t}$ being the time instants related to the frame, this term is equal to $\ln p_{x_b}\left(x - 2\text{Re}\left(\sum_{p=1}^{P}\sum_{h=1}^{H_p}\alpha_{hp}e^{2i\pi f_{hp}\mathbf{t}}\right)|\theta_b\right)$



(a) True note: A♭4. Selected model: A♭4.

(b) True note: A♭4. Selected model: A♭3.

Fig. 2. Estimation of spectral envelopes: the AR estimation performs a better fitting of the amplitudes for the correct model (left) than for the sub-octave model (right).

As c is a constant w.r.t. A_p, the optimization consists in maximizing $\rho(A_p)$, which measures the spectral flatness (*i.e.* the ratio between the geometrical and arithmetical means) of $\left\{|\alpha_{hp}|^2 \left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2\right\}_{1\leq h\leq H_p}$. This quantity lies in $[0;1]$ and is maximum when the latter coefficients are constant, *i.e.* when the filter A_p perfectly whitens the amplitudes α_p. The estimation of A_p from the discrete data α_p thanks to the Digital All-Pole (DAP) method [22] leads to a solution that actually maximizes $\rho(A_p)$, as proved in [23]. It is applied here to the amplitudes α_p, for each note p.

The estimation of the spectral envelope of a note is illustrated in Fig. 2. A note is analyzed by considering two possible models: the model related to the true note (Fig. 2(a)) and the model related to its sub-octave (Fig. 2(b)). In the former case, the estimate of the spectral envelope is close to the amplitudes of the overtones, whereas in the latter case, the AR spectral envelope model is not adapted to amplitudes that are alternatively high and low: the low values obtained for the spectral flatness ρ (and, consequently, for the likelihood) are here a good criterion to reject wrong models like the sub-octave (see Fig. 6(a) for the whole spectral flatness curve).

2) Estimation of the amplitudes of the overtones: let us now assume that spectral envelope parameters θ are known, that the frequencies of the overtones may overlap, and that the amplitudes α are unknown. In all that follows, we assume that at a given overtone frequency f_{hp}, the power spectrum of the noise is not significant in comparison with the power spectrum of the amplitudes of the overtone.

When the overtone h of note p is not overlapping with any other overtone, the amplitude α_{hp} is directly given by the spectrum value $X(f_{hp})$. In the alternative case of overlapping overtones, the observed spectrum results from the contribution of overtones with close frequencies. We propose an estimate of the hidden random variable α, given the observation X and the parameters θ that control the second-order statistics of α.

As defined by eq. (3), the amplitude α_{hp} is a random variable such that $\alpha_{hp} \sim \mathcal{N}(0, v_{hp})$ with $v_{hp} \triangleq \frac{\sigma_p^2}{\left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2}$. One can rewrite the sound model x as a sum of K sinusoids and noise $x(t) = \sum_{k=1}^{K} \alpha_k e^{2i\pi f_k t} + x_b(t)$ with $\alpha_k \sim \mathcal{N}(0, v_k)$ and $K = 2\sum_{p=1}^{P} H_p$. In this formula, the couples of indexes (h,p) related to overtones and notes have been replaced by a single index k. In the spectral domain, the DFT X of the N-length frame x using a weighting window w of DFT W is $X(f) = \sum_{k=1}^{K} \alpha_k W(f - f_k)$, the noise spectrum being removed since we are only interested in the values of X at frequencies f_{hp}, where the noise term is insignificant.

In these conditions, for $1 \leq k_0 \leq K$, the optimal linear estimator $\hat{\alpha}_{k_0}$ of α_{k_0} as a function of $X(f_{k_0})$, obtained by minimizing the mean squared error $\epsilon_{k_0} = \mathbb{E}\left[\left|\alpha_{k_0} - \hat{\alpha}_{k_0}\right|^2\right]$ is

$$\hat{\alpha}_{k_0} = \frac{W^*(0)\, v_{k_0}}{\sum_{k=1}^{K} |W(f_{k_0} - f_k)|^2 v_k} X(f_{k_0}) \qquad (12)$$

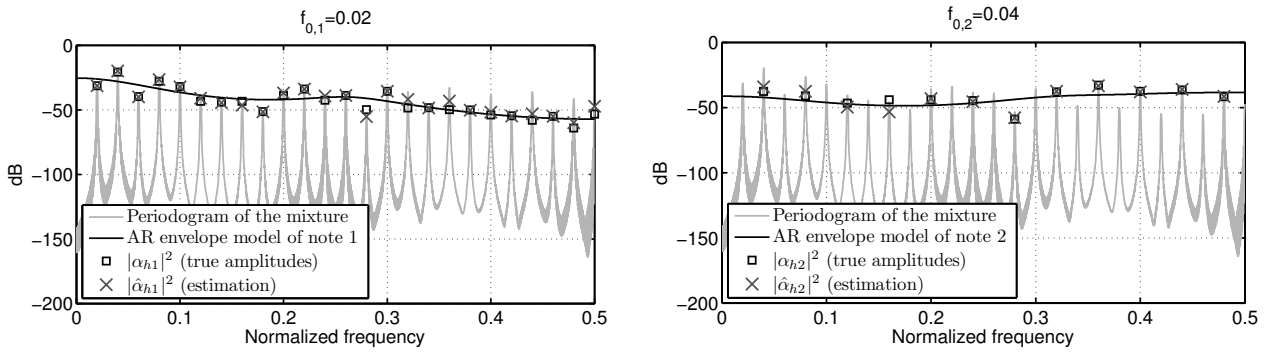
The proof of eq. (12) is given in the Appendix B. By ignoring non-overlapping overtones, the expression of the estimator of α_{hp} can be simplified as

$$\widehat{\alpha_{hp}} \triangleq \frac{W^*(0)\, v_{hp}^2}{\sum\limits_{\left|f_{h'p'} - f_{hp}\right| < \Delta_w} \left|W\left(f_{h'p'} - f_{hp}\right)\right|^2 v_{h'p'}^2} X(f_{hp}) \qquad (13)$$

where Δ_w is the width of the main lobe of W.

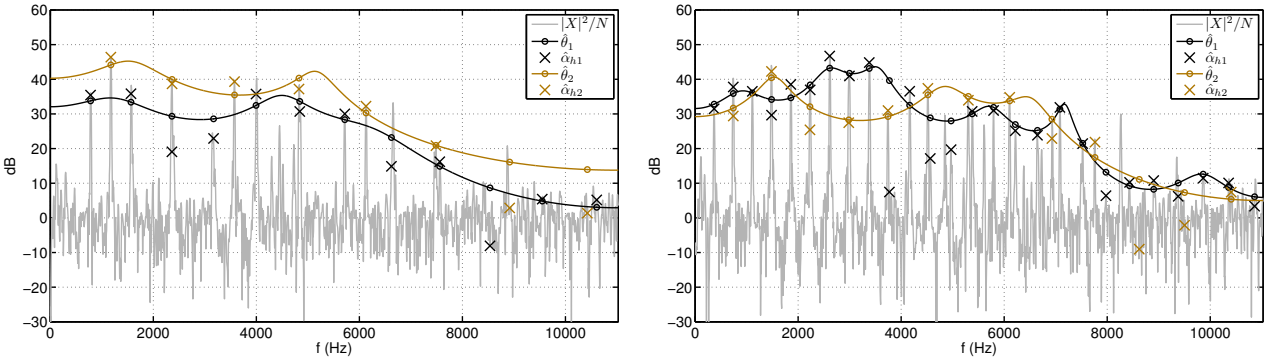
The amplitude estimation is illustrated in Fig. 3 on a synthetic signal, composed of two octave-related notes, *i.e.* with overlapping spectra. As expected, the amplitudes are perfectly estimated when there is no overlap (see odd-order peaks of note 1). When the overtones overlap, predominant amplitudes are well estimated (see note 1 for $f = 0.04$ or note 2 for $f = 0.4$), as well as amplitudes with the same order of magnitude (see amplitudes at $f = 0.2$). The estimation may be less accurate when an amplitude is much weaker than the other (see note 1 at $f = 0.36$).

3) Iterative algorithm for the estimation of note parameters: in the last two parts, we have successively seen how to estimate the spectral envelope models from known amplitudes and the amplitudes from known spectral envelope models. None of these quantities are actually known, and both must be estimated. We propose to use the two methods above in the Algorithm 1 to iteratively estimate both the spectral envelope and the amplitudes of the overtones. It alternates the estimation of the former and of the latter, with an additional thresholding of the weak amplitudes in order to avoid the amplitude estimates becoming much lower than the observed



(a) Note 1. (b) Note 2.

Fig. 3. Estimation of amplitudes from overlapping spectra: example on a synthetic mixture of 2 notes. The F_0s $f_{0,1}$ and $f_{0,2}$ of the two notes are in an octave relation and the estimation is performed from a $N = 2048$-length observation. For each note, the true amplitudes (squares) have been generated from a spectral envelope AR model (black line) using eq. (3). The amplitudes are estimated (crosses) from the observation of the mixture (grey line) using the AR envelope information (eq. (12)).



(a) True notes: G_5+D_6. Selected model: G_5+D_6. (b) True notes: F#$_4$+F#$_5$. Selected model: F#$_4$+F#$_5$.

Fig. 4. Iterative estimation of amplitudes $\widehat{\alpha_p}$ and spectral envelopes $\widehat{\theta_p}$: example of a fifth (Fig. 4(a)) and of an octave (Fig. 4(b)).

spectral coefficients[5]. For $p \in [\![1; P]\!]$ and $h \in [\![1; H_p]\!]$, the estimate $\widehat{\alpha_{hp}}$ of α_{hp} and $\widehat{\theta_p}$ of θ_p are respectively the values of $\left(\alpha_{hp}^{(i)}\right)$ and $\left(\theta_p^{(i)}\right)$ after a number of iterations. Note that the ML estimates of the envelopes are directly related to the MAP estimation of the multipitch contents while amplitudes are estimated in the least mean square sense, which is a different objective function. Consequently, the convergence of the iterative algorithm is not proved or guaranteed but it is observed after about $N_{\text{it}} \triangleq 20$ iterations.

The estimation is illustrated in Fig. 4 with two typical cases. While the energy is split between the overlapping overtones, non-overlapping overtones help estimating the spectral envelopes, resulting in a good estimation in the case of a fifth (Fig. 4(a)), the octave being a more difficult – but successful – case (Fig. 4(b)).

Algorithm 1 Iterative estimation of note parameters

Require: spectrum X, notes $\mathcal{C}_1, \ldots, \mathcal{C}_P$.
  Initialize the amplitudes $\alpha_{hp}^{(0)}$ of overtones to spectral values $X(f_{hp})$ for $p \in [\![1; P]\!]$ and $h \in [\![1; H_p]\!]$.
  for each iteration i **do**
    for each note p **do**
      Estimate $\theta_p^{(i)}$ from $\alpha_p^{(i-1)}$. {AR estimation}
    end for
    Jointly estimate all $\alpha_{hp}^{(i)}$ from $X(f_{hp})$ and $\theta_p^{(i)}$. {eq. (13)}
    Threshold $\alpha_{hp}^{(i)}$ at a minimum value.
  end for
Ensure: estimation of α_{hp} and θ_p for $p \in [\![1; P]\!]$ and $h \in [\![1; H_p]\!]$.

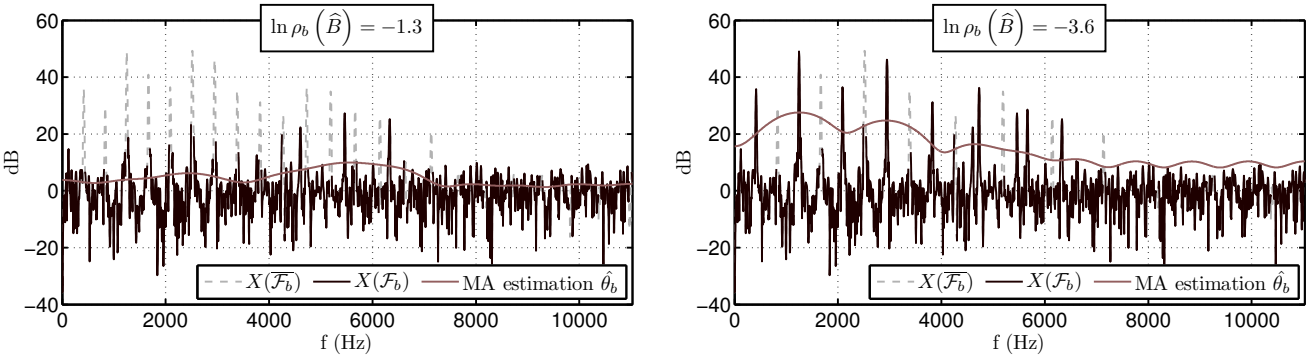
B. Estimation of noise parameters

We assume that the noise signal results from the circular filtering[6] of a white centered Gaussian noise with variance σ_b^2 by a Q_b-order FIR filter parameterized by its transfer function $B(z) = \sum_{k=0}^{Q_b} b_k z^{-k}$, with $b_0 = 1$.

As we assumed that in a bin close to the frequency of an overtone, the spectral coefficients of noise are negligible w.r.t. the spectral coefficients of notes, the information related to noise is mainly observable in the remaining bins, *i.e.* on the

[5]In our implementation, this threshold was set to the minimum observed amplitude $\min_f \frac{|X(f)|}{\sqrt{N}}$.

[6]The circularity assumption, which is commonly used and asymptotically valid, leads to a simplified ML solution.



(a) True note: $E\flat_4$. Selected model: $E\flat_4$.

(b) True note: $E\flat_4$. Selected model: $E\flat_5$.

Fig. 5. Noise parameter estimation when selecting the true model (left) and the octave model (right). $\overline{\mathcal{F}_b}$ denotes the complement of $\mathcal{F}_b$.

frequency support defined by

$$\mathcal{F}_b \triangleq \left\{ \frac{k}{N_{\text{fft}}} \middle/ \forall p \in [\![1;P]\!], \forall h \in [\![1;H_p]\!], \left| \frac{k}{N_{\text{fft}}} - f_{hp} \right| > \frac{\Delta_w}{2} \right\} \tag{14}$$

where Δ_w denotes the width of the main lobe of w ($\Delta_w = \frac{4}{N}$ for a Hann window) and N_{fft} is the size of the DFT. The noise parameter estimation is then performed using this subset of bins, which gives satisfying results asymptotically (*i.e.* when $N \to +\infty$, the number of removed bins being much smaller than the total number of bins). As proved in the Appendix B, it results in the maximization, w.r.t. B, of the expression

$$L_b\left(\widehat{\sigma_b^2}, B\right) = c_b + \frac{|\mathcal{F}_b|}{2} \ln \rho_b\left(B\right) \tag{15}$$

where

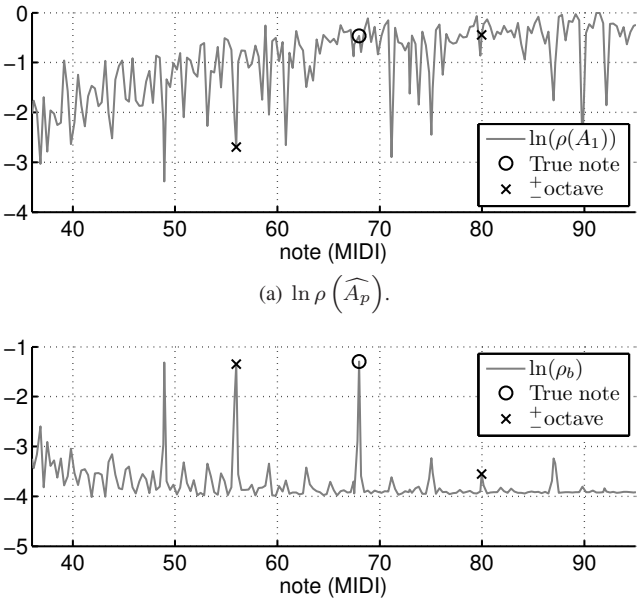
$$c_b \triangleq -\frac{|\mathcal{F}_b|}{2} \ln 2\pi e - \frac{1}{2} \sum_{f \in \mathcal{F}_b} \ln \frac{\left|X_b\left(f\right)\right|^2}{N} \tag{16}$$

$$\rho_b\left(B\right) \triangleq \frac{\left(\prod_{f \in \mathcal{F}_b} \left| \frac{X_b(f)}{B(e^{2i\pi f})} \right|^2 \right)^{\frac{1}{|\mathcal{F}_b|}}}{\frac{1}{|\mathcal{F}_b|} \sum_{f \in \mathcal{F}_b} \left| \frac{X_b(f)}{B(e^{2i\pi f})} \right|^2} \tag{17}$$

$$\widehat{\sigma_b^2} = \frac{1}{|\mathcal{F}_b|} \sum_{f \in \mathcal{F}_b} \frac{1}{N} \left| \frac{X_b\left(f\right)}{B\left(e^{2i\pi f}\right)} \right|^2 \tag{18}$$

As for the spectral envelope AR parameters (eq. (11) and (8)), the ML estimation of noise parameters consists in the maximization of a spectral flatness $\rho_b\left(B\right)$. It can be achieved thanks to the MA estimation approach proposed in [23]. However, in order to speed up the computations, we use the Algorithm 3 described in the Appendix C, which is simpler and gives satisfying results for the current application.

The noise parameter estimation is illustrated in Fig. 5. When the true model is selected (Fig. 5(a)), most of primary lobes of the sinusoidal components are removed from $\mathcal{F}_b$ and the resulting spectral coefficients are well-fitted by an MA envelope. In the case a wrong model is selected and a lot of $\mathcal{F}_b$ bins are related to sinusoids (Fig. 5(b)), the MA model is not adapted to the observations which are not well-fitted. The resulting ρ_b value and likelihood will be low so that the wrong model will be rejected. As shown in Fig. 6, the spectral flatness criterion for the spectral envelope and noise estimations are complementary cues for rejecting wrong models and selecting the right one. The criterion on spectral envelopes (Fig. 6(a)) shows many high values but also large minima for sub-harmonic errors (*e.g.* notes $D\flat_3$, $A\flat_3$ and $D\flat_3$ in this example). Conversely, the criterion for the noise model (Fig. 6(b)) has a few high peaks located at the sub-harmonics F_0s of the true note – *i.e.* for every model all the sinusoids have been removed from the noise observation –, but is efficient for discriminating the other errors. Thus, if taken separately, these criteria are not good pitch estimators, but they efficiently combine into the detection function.



(a) $\ln \rho\left(\widehat{A_p}\right)$.

(b) $\ln \rho_b\left(\widehat{B}\right)$.

Fig. 6. $\ln \rho\left(\widehat{A_1}\right)$ (top) and $\ln \rho_b\left(\widehat{B}\right)$ (bottom) as a function of a single-note model (true note: $A\flat_4$).

IV. Experimental results

A. Implementation and tuning

The whole method is summarized by Algorithm 2. The MPE algorithm is implemented in Matlab and C. It is designed to analyze a 93ms frame and to estimate its polyphonic content. A preprocessing stage aims at reducing the spectral dynamics and consists in flattening the global spectral decrease by means of a median filtering of the spectrum. The values of the parameters of the MPE algorithm are given in Table I. The order Q_p of the spectral envelope filter $A_p(z)$ of a note p is set to $Q_p = H_p/2$ in order to be large enough to fit the data and small enough to obtain a smooth spectral envelope[7]. Parameters $\left(k_{\sigma_p^2}, E_{\sigma_p^2}\right)$, $\left(k_{\sigma_b^2}, E_{\sigma_b^2}\right)$, $\mu_{\text{env}}, \mu_{\text{b}}, \mu_{\text{pol}}$ and $(w_1, \ldots, w_4)$ have been estimated on a development database composed of about 380 mixtures from two different pianos, with polyphony levels from 1 to 6. They were jointly adjusted by optimizing the F-measure using a grid search, the bounds of the grid being set by hand.

Algorithm 2 Multipitch estimation.

Preprocessing
Computation of periodogram $\frac{1}{N}|X|^2$
Selection of N_c note candidates
for each possible note combination $\mathcal{C}$ **do**
  Iterative estimation of the amplitudes of overtones and of spectral envelope of notes (algorithm 1)
  For each note p, derivation of L_p (eq. (8)) and of prior $p\left(\widehat{\sigma_p^2}\right)$
  Noise parameter estimation (algorithm 3)
  Computation of L_b (eq. (15)) and of prior $p\left(\widehat{\sigma_b^2}\right)$
  Derivation of the detection function related to the note combination (eq. (7))
end for
Maximization of the detection function

Parameter	value	Parameter	value
$P_{\max}$	6	$\left(k_{\sigma_p^2}, E_{\sigma_p^2}\right)$	$\left(1, 10^{-4}\right)$
N_c	9	$\left(k_{\sigma_b^2}, E_{\sigma_b^2}\right)$	$\left(2, 10^{-3}\right)$
ν	0.38	μ_{env}	$-8.9{\cdot}10^{-3}$
f_s	22050Hz	μ_b	$-2.2{\cdot}10^{-4}$
N	2048	μ_{pol}	25
w	Hann	w_1	$8.1{\cdot}10^{-1}$
Q_p	$H_p/2$	w_2	$1.4{\cdot}10^{4}$
Q_b	20	w_3	$6.2{\cdot}10^{2}$
N_{it}	20	w_4	5.8

TABLE I
Parameters of the algorithm.

B. Evaluation

The proposed MPE algorithm has been tested on a database called MAPS[8] and composed of around 10000 piano sounds either recorded by using an upright Disklavier piano or generated by several virtual piano software products based on sampled sounds. The development set and the test set are disjointed. Two pianos are used in the former while the latter comprises sounds from other five pianos. In total, two upright pianos and five grand pianos were used. Recording conditions and tunings vary from one instrument to the other. Polyphony levels lie between 1 and 6 and notes are uniformly distributed between C_2 (65Hz) and B_6 (1976Hz). One part of the polyphonic mixtures is composed of randomly related pitches whereas the other part comprises usual chords from western music (major, minor, etc.). For each sound, a single 93ms-frame located 10ms after the onset time is extracted and analyzed. Two additional algorithms have been tested on the same database for comparison purposes. The first one is Tolonen's multipitch estimator [24] for which the implementation from the MIR Toolbox [25] was used. As the performance of the algorithm decreases when F_0s are greater than 500Hz (C_5), the system was additionally tested in restricted conditions – denoted Tolonen-500 – by selecting from the database the subset of sounds composed of notes between C_2 and B_4 only. The second one is Klapuri's system [26]. The code of the latter was provided by its author.

Our algorithm was implemented in Matlab and C. The computational cost on a recent PC is about $150\times$real time. It thus requires more computations than other algorithms like [26] but the proposed algorithm is computationally tractable, particularly when comparing it to the greedy and intractable joint-estimation approach where no note candidate selection is performed, as discussed in section II-B.

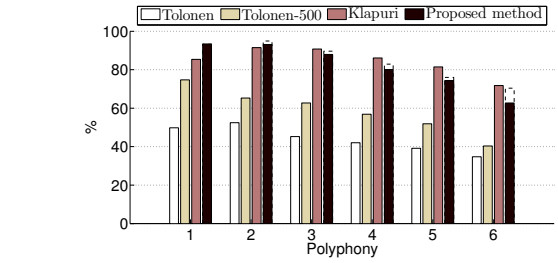
General results are presented in Fig. 7. Relevant items are defined as correct notes after rounding each F_0 to the nearest half-tone. Typical metrics are used: the recall is the ratio between the number of relevant items and of original items; the precision is the ratio between the number of relevant items and of detected items; and the F-measure is the harmonic mean between the precision and the recall. In this context, our system performs the best results for polyphony 1 and 2: 93% F-measure vs 85% (polyphony 1) and 91% (polyphony 2) for Klapuri's system. The trend then reverses between Klapuri's system and the proposed one, the F-measure being respectively 91% and 88% for polyphony 3, and 72% and 63% for polyphony 6. Moreover, the precision is high for all polyphony levels whereas the recall is decreasing when polyphony increases. Results from Tolonen's system are weaker for the overall tests, even with the restricted F_0-range configuration.

The ability of Klapuri's system and ours to detect polyphony levels (independently of the pitches) is presented in Tab. II. For polyphony levels from 1 to 5, both systems succeed in detecting the correct polyphony level more often than any other level. It should also be noted that the proposed system was tested in more difficult conditions since polyphony 0 (*i.e.* silence) may be detected, which is the case for a few sounds[9]. The proposed system tends to underestimate the polyphony

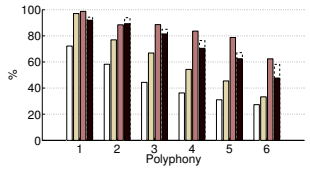
[7]Note that the number of degrees of freedom for the AR models is $H_p/4 = Q_p$ since half the poles are affected to the positive frequencies, the remaining poles being their complex conjugates.

[8]MAPS stands for MIDI Aligned Piano Sounds and is available on request.

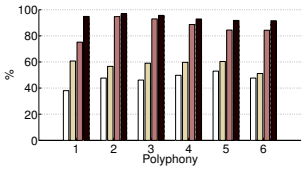
[9]Note that silence detection could be improved by using an activity detector as a preprocessing stage, since silence is a very specific case of signal that can be detected using more specific methods than the proposed one.



(a) F-measure



(b) Recall



(c) Precision

Fig. 7. Multipitch estimation results with unknown polyphony: for each algorithm, the F-measure, the precision and the recall are rendered as a function of the true polyphony. For the proposed method, the performance on the training set is plotted using a dashed line.

	P					
P_{est}	1	2	3	4	5	6
0	7	2	4	6	9	0
1	**90**	14	8	8	8	12
2	3	**83**	18	12	8	19
3	0	1	**68**	27	16	25
4	0	0	2	**47**	27	32
5	0	0	0	1	**31**	11
6	0	0	0	0	1	**1**

Proposed method

	P					
P_{est}	1	2	3	4	5	6
0	0	0	0	0	0	0
1	**92**	21	5	2	1	3
2	1	**74**	18	7	4	7
3	1	3	**65**	24	10	12
4	0	1	9	**49**	24	27
5	1	0	1	14	**36**	23
6	5	1	1	4	25	**28**

Klapuri's method

TABLE II
POLYPHONY ESTIMATION: DETECTION RATE (%) W.R.T. TO THE NUMBER OF SOUNDS WITH TRUE POLYPHONY P, AS A FUNCTION OF THE TRUE POLYPHONY P AND OF THE ESTIMATED POLYPHONY P_{EST}.

level since the parameter tuning consists in optimizing the F-measure on the development set. This objective function could have been changed to take the polyphony level balance into account. This would result in reducing the polyphony underestimation trend. However, the overall F-measure would decrease. From a perceptive point of view, it has been shown that a missing note is generally less annoying than an added note when listening to a resynthesized transcription [27]. Thus, underestimating the polyphony may be preferred to overestimating it. Still, this trend turns out to be the main shortcoming of the proposed method, and should be fixed in the future for efficiently addressing sounds with polyphony higher than 5 notes.

The case of octave detection has been analyzed and results are presented in Fig. 8. Octave detection is one of the most difficult cases of multipitch estimation and the results are lower than the global results obtained for polyphony 2 (see Fig. 7). The proposed method reaches the highest results here, the F-measure being 81%, versus 77% for Klapuri's system and $77\%/66\%$ for Tolonen's (depending on the two F_0 ranges). This would suggest that the proposed models for spectral envelopes and overlapping spectra are significatively efficient.

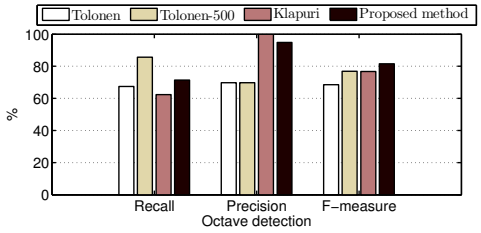


Fig. 8. Octave detection: results for the 97 octave sounds extracted from the database (45 sounds for the Tolonen-500 system).

Finally, we report additional results which are not depicted. First, if polyphony is known, the performance of our system reaches the second rank after Klapuri's system, with correct detection rates from 95% (polyphony 1) down to 55% (polyphony 6). Second, F-measure values are between 5% and 10% better for usual chords than for random-pitch chords. This has been observed with all the tested methods for polyphony levels higher than two. While the algorithms face more harmonically-related notes in usual chords – *i.e.* more spectral overlap –, it seems that simultaneous notes with widely-spread F_0s in random chords are a bigger difficulty. Third, the results obtained on the development set and the test set are comparable, with only a few $\%$ deviation. This shows that the parameter learning is not overfitting and that the system has good generalization abilities for other models of piano and different recording conditions. Fourth, the whole database includes sounds from seven different pianos. The results are comparable from one piano to the other, with about 3%-standard deviation. Results do not significantly depend on the upright/grand piano differences, on recording conditions or on whether a real piano or a software-based one is used, which suggests robustness for varying production conditions. Fifth, the proposed method uses a candidate selection stage which may fail, causing the subsequent multipitch estimation to fail. The performance of the candidate selection stage has thus been checked: satisfying results are obtained for low to medium polyphony levels, 99% of true notes being selected for polyphony 1 and 2, 94% for polyphony 3 and 86% for polyphony 4. Performance is slightly decreasing in polyphony 5 and 6, the scores being 78% and 71% respectively. Along the original-pitch dimension, these errors are mainly located below A_2 (110Hz), and above A_6 (1760Hz). Selecting a set of likely F_0s is a challenging issue for future works since a lot of MPE algorithms would benefit from it. Indeed, they often consist in optimizing an objective criterion that presents a lot of local optima along the pitch dimension. Some basic solutions like exhaustive or grid search [27] or more elaborated ones like Markov chain Monte Carlo methods [28] have already been proposed for searching the global optimum, and an efficient candidate selection stage would be very helpful to reach the result more quickly.

C. Integration in an automatic transcription system

The proposed MPE method has been introduced for a single-frame analysis. Beyond the multipitch issue, the automatic transcription task is addressed by integrating this method in

a system that performs pitch tracking over time and outputs not only pitches but note events. In the current case, the MPE method can be used in the transcription framework proposed in [19]. The automatic transcription of piano music is thus obtained by:

1) using an onset detector to localize new events;
2) detecting a set of pitch candidates after each onset;
3) between two consecutive onsets, decoding the multipitch contents thanks to a frame-based Hidden Markov Model (HMM), the states being the possible combinations of simultaneous pitches, the likelihood being given by the proposed MPE method (eq. (7));
4) postprocessing the decoded pitches to detect repetitions or continuations of pitches when an onset occurs.

This transcription system was presented at the *Multiple Fundamental Frequency Estimation & Tracking* task of the MIREX[10] contest in 2008 (system called EBD2). For the piano-only test set, the evaluation is performed in two different ways, depending on how a correct note is defined. In the first, more constraining evaluation, where a correct note implies a correct onset (up to a 50-ms deviation), a correct offset (up to a 20%-duration or 50-ms deviation) and a correct pitch (up to a quartertone deviation), the proposed system reaches the 3rd rank with a 33.1% F-measure, out of 13 systems scoring from 6.1% to 36.8%. For that evaluation, the duration estimation is evaluated by the average overlap between correct estimations and reference notes: the proposed method reaches the 7th position with 80%, the values being in a narrow interval ([77.4%; 84%]). In the second evaluation, where correctness only requires the onset and pitch criteria above, the proposed system reaches the 7th rank with a 56.9% F-measure, out of 13 systems scoring from 24.5% to 75.7%. The best rank is obtained for average overlap with a 61% value, the lowest value being 40.1%.

Thus, when the proposed MPE method is integrated in a full transcription system for piano music, state-of-the art results are reached in terms of global performance. In addition, good average overlap scores show that the proposed method is suitable for pitch tracking over time and note extraction.

Note that one can wonder whether the reported performance is significant when comparing MPE and ATM quantitative results (with roughly 20% F-measure deviation, no matter the method is). It actually depends not only on the database, but on the different testing protocols. While isolated frames composed of one chord are used in the proposed MPE evaluation, ATM evaluation implies other difficulties like having asynchronous notes overlaping in time; detecting onsets; estimating the end of damping notes; dealing with reverberation queues and so on. More, polyphony levels in musical recordings are often high and are not uniformly distributed at all, with a 4.5 average polyphony and a 3.1 standard deviation reported for a number of classical music pieces [29, p.114]. Hence, F-measure for MPE and ATM should be evaluated separately.

[10]http://www.music-ir.org/mirex/2008/

V. Conclusions

In this paper, a sound model and a method were proposed for the multipitch estimation problem in the case of piano sounds. The approach was based on an adaptive scheme, including the adaptive matching of the inharmonic distribution of the frequencies of the overtones and an autoregressive spectral envelope. We have shown the advantage of using a moving average model for the residual noise. The estimation of the parameters was performed by taking the possible overlap between the spectra of the notes into account. Finally, a weighted likelihood criterion was introduced to determine the estimated mixture of notes in the analyzed frame among a set of possible mixtures.

The performance of the method was measured on a large database of piano sounds and compared to the state-of-the-art. The proposed method provides satisfying results when polyphony is unknown, and reaches particularly good scores for mixtures of harmonically-related notes.

This approach has been successively integrated in a full transcription system [19] and alternative investigations [30] are ongoing. Future works may deal with the use of different spectral envelopes in a similar modeling and estimation framework. Indeed, while the proposed AR envelope model may be too generic to characterize the spectral envelope smoothness of musical instruments, this information can be easily replaced, as the variance of the amplitude Gaussian models, by any parametric envelope model. The method may also be extended to other instruments or to voice. Keeping the proposed sound model, it could be useful to investigate other estimation strategies in order to reduce the computational cost. This may be achieved by means of a hybrid approach with an iterative selection of the notes and a joint estimation criterion. Furthermore, the log-likelihood function related to the model for spectral envelopes is not very selective. Hence, some investigations on more appropriate functions could improve the performance of decision step. Finally, one could conduct others investigations about how to prune the set of all possible note combinations more efficiently. They may include iterative candidate selection or MCMC methods.

Appendix A
Discussion on the detection function

According to eq. (5), the multipitch estimation is theoretically obtained by maximizing, w.r.t. $\mathcal{C}$, the likelihood $p(x|\mathcal{C})$, which is expressed as a function of the parameters of the proposed model as:

$$p(x|\mathcal{C}) = \iint p(x,\alpha,\theta_b|\mathcal{C})\, d\alpha d\theta_b \tag{19}$$

$$= \iint p(x|\alpha,\theta_b,\mathcal{C})\, p(\alpha|\mathcal{C})\, p(\theta_b)\, d\alpha d\theta_b \tag{20}$$

$$= \iiint p(x|\alpha,\theta_b,\mathcal{C})\, p(\alpha|\theta,\mathcal{C})\, p(\theta)\, p(\theta_b)\, d\theta d\alpha d\theta_b \tag{21}$$

Since the derivation of eq. (21) is not achievable, a simpler detection function must be built. The proposed function can be interpreted as a weighted log-likelihood and is obtained from the integrand of eq. (21) by:

- considering the estimated parameters $\widehat{\theta_p}, \widehat{\alpha_p}, \widehat{\theta_b}$;
- taking the logarithm, thus turning the product into a sum;
- normalizing the sums over the number of notes by P, and log-likelihoods L_p and L_b by the size H_p and $|\mathcal{F}_b|$ of their respective random variable;
- introducing penalization terms using the $\mathcal{C}$-dependent sizes P, H_p, $|\mathcal{F}_b|$ with coefficients $\mu_{\text{pol}}, \mu_{\text{env}}, \mu_b$;
- weighting each quantity by the coefficients $w_1, \ldots, w_4$.

The normalization operation aims at obtaining data that can be compared from one mixture $\mathcal{C}$ to another, since the dimension of the random variables depends on the mixture. The penalization is inspired by model selection approaches [31], in which an integral like (21) is replaced by a likelihood penalized by a linear function of the model order. The weighting by $w_1, \ldots, w_4$ is finally used to find an optimal balance between the various probability density functions. Note that in eq. (7), priors related to filters A_p and B have been removed since they are non-informative, and thus can be considered as constant terms.

APPENDIX B
PROOFS

Proof of eq. (8)*:* Using eq. (3), the log-likelihood to be maximized is

$$L_p\left(\sigma_p^2, A_p\right) = -\frac{H_p}{2}\ln\left(2\pi\sigma_p^2\right) + \frac{1}{2}\sum_{h=1}^{H_p}\ln\left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2 - \frac{1}{2\sigma_p^2}\sum_{h=1}^{H_p}\left|\alpha_{hp}\right|^2\left|A_p\left(e^{2i\pi f_{hp}}\right)\right|^2 \tag{22}$$

The maximization w.r.t. σ_p^2 leads to the estimate $\widehat{\sigma_p^2}$ given by eq. (11), which is used in eq. (22) to obtain eq. (8). ∎

Proof of eq. (12) *and* (13)*:* The expected linear estimator is expressed as $\hat{\alpha}_{k_0} \triangleq \eta X\left(f_{k_0}\right)$ with $\eta \in \mathbb{C}$. The mean squared error is then $\epsilon_{k_0}(\eta) = \mathbb{E}\left[\left|\alpha_{k_0} - \eta X\left(f_{k_0}\right)\right|^2\right]$. The optimal value $\hat{\eta}$ is such that $\frac{d\epsilon_{k_0}}{d\eta}(\hat{\eta}) = 0$, which is equivalent to the decorrelation between the error $\left(\alpha_{k_0} - \hat{\eta}X\left(f_{k_0}\right)\right)$ and the data $X\left(f_{k_0}\right)$:

$$\begin{aligned} 0 &= \mathbb{E}\left[\left(\alpha_{k_0}^* - \hat{\eta}^* X^*\left(f_{k_0}\right)\right) X\left(f_{k_0}\right)\right] \\ &= \mathbb{E}\left[\alpha_{k_0}^* X\left(f_{k_0}\right)\right] - \hat{\eta}^*\mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right] \end{aligned} \tag{23}$$

which leads to $\hat{\eta} = \frac{\mathbb{E}\left[\alpha_{k_0} X^*\left(f_{k_0}\right)\right]}{\mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right]}$, where

$$\begin{aligned} \mathbb{E}\left[\alpha_{k_0} X^*\left(f_{k_0}\right)\right] &= \sum_{k=1}^{K} W^*\left(f_{k_0} - f_k\right)\mathbb{E}\left[\alpha_{k_0}\alpha_k^*\right] \\ &= W^*(0)\, v_{k_0} \end{aligned} \tag{24}$$

$$\begin{aligned} \text{and } \mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right] &= \sum_{k=1}^{K}\sum_{l=1}^{K} W\left(f_{k_0} - f_k\right) \\ &\quad \times W^*\left(f_{k_0} - f_l\right)\mathbb{E}\left[\alpha_k\alpha_l^*\right] \\ &= \sum_{k=1}^{K}\left|W\left(f_{k_0} - f_k\right)\right|^2 v_k \end{aligned} \tag{25}$$

The related error is

$$\begin{aligned} \epsilon_{k_0}(\hat{\eta}) &= \mathbb{E}\left[\left|\alpha_{k_0}\right|^2\right] + \left|\hat{\eta}\right|^2\mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right] \\ &\quad - \hat{\eta}\mathbb{E}\left[\alpha_{k_0}^* X\left(f_{k_0}\right)\right] - \hat{\eta}^*\mathbb{E}\left[\alpha_{k_0} X^*\left(f_{k_0}\right)\right] \\ &= \mathbb{E}\left[\left|\alpha_{k_0}\right|^2\right] - \frac{\left|\mathbb{E}\left[\alpha_{k_0} X^*\left(f_{k_0}\right)\right]\right|^2}{\mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right]} \\ &= \left(1 - \frac{\left|W(0)\right|^2 v_{k_0}}{\sum_{k=1}^{K}\left|W\left(f_{k_0} - f_k\right)\right|^2 v_k}\right) v_{k_0} \end{aligned} \tag{26}$$

Note that this approach has some connections with the Wiener filtering technique [32] widely used in the field of audio source separation. In both cases, the aim is to estimate a hidden variable thanks to its second-order statistics and to the observations. In the considered short-term analysis context, the proposed approach explicitly models the overlap phenomenon and takes the resulting frequency leakage into account. As expected, the estimation error is all the larger as the estimated overtone is masked by another component, which justifies the approximation of eq. (12) in eq. (13). ∎

Proof of eq. (15)*:* In a matrix form, the filtering process is expressed as $\underline{x_b} = B_{\text{circ}}\underline{w_b}$, where $\underline{x_b}$ is an N-length frame of the noise process x_b, $\underline{w_b} \sim \mathcal{N}\left(0, \boldsymbol{I}_N\sigma_b^2\right)$ and B_{circ} is the $N \times N$ circulant matrix with first column $(b_0, \ldots, b_{Q_b}, 0, \ldots, 0)$. Thus, we have $\underline{x_b} \sim \mathcal{N}\left(0, B_{\text{circ}}\left(B_{\text{circ}}\right)^\dagger\sigma_b^2\right)$. B_{circ} is circulant, so $\det\left(B_{\text{circ}}B_{\text{circ}}^\dagger\right) = \prod_{k=0}^{N-1}\left|B\left(e^{2i\pi\frac{k}{N}}\right)\right|^2$. The likelihood of $\underline{x_b}$ is then

$$p\left(\underline{x_b}\right) = \frac{e^{-\frac{1}{2}\underline{x_b}^\dagger\left(B_{\text{circ}}B_{\text{circ}}^\dagger\sigma_b^2\right)^{-1}\underline{x_b}}}{\sqrt{(2\pi)^N\det\left(B_{\text{circ}}B_{\text{circ}}^\dagger\sigma_b^2\right)}} \tag{27}$$

$$= \frac{e^{-\frac{\left\|B_{\text{circ}}^{-1}\underline{x_b}\right\|^2}{2\sigma_b^2}}}{\sqrt{\left(2\pi\sigma_b^2\right)^N\prod_{k=0}^{N-1}\left|B\left(e^{2i\pi\frac{k}{n}}\right)\right|^2}} \tag{28}$$

In the spectral domain, the log-likelihood of the noise signal

is thus obtained using the Parseval identity:

$$L_b\left(\sigma_b^2, B\right) = -\frac{N}{2}\ln 2\pi\sigma_b^2 - \frac{1}{2}\sum_{k=0}^{N-1}\ln\left|B\left(e^{2i\pi\frac{k}{N}}\right)\right|^2 - \frac{1}{2\sigma_b^2}\sum_{k=0}^{N-1}\frac{1}{N}\left|\frac{X_b\left(\frac{k}{N}\right)}{B\left(e^{2i\pi\frac{k}{N}}\right)}\right|^2 \quad (29)$$

When only considering the observations on the spectral bins $\mathcal{F}_b$, the log-likelihood (29) becomes

$$L_b\left(\sigma_b^2, B\right) = -\frac{|\mathcal{F}_b|}{2}\ln 2\pi\sigma_b^2 - \frac{1}{2}\sum_{f\in\mathcal{F}_b}\ln\left|B\left(e^{2i\pi f}\right)\right|^2 - \frac{1}{2\sigma_b^2}\sum_{f\in\mathcal{F}_b}\frac{1}{N}\left|\frac{X_b\left(f\right)}{B\left(e^{2i\pi f}\right)}\right|^2 \quad (30)$$

The maximization w.r.t. σ_b^2 leads to the estimate given by eq. (18), which is used in eq. (30) to obtain eq. (15).

■

APPENDIX C
MA ESTIMATION

The algorithm is based on the decomposition $\underline{r_b} = \mathbf{B}\underline{b}$ of the first $(Q_b + 1)$ terms $\underline{r_b}$ of the autocorrelation function of the process, as a product of the coefficient vector $\underline{b} \triangleq (b_0, \ldots, b_{Q_b})^{\mathrm{T}}$ by the matrix

$$\mathbf{B} \triangleq \begin{pmatrix} b_0 & b_1 & \ldots & b_{Q_b} \\ 0 & b_0 & \ldots & b_{Q_b-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & b_0 \end{pmatrix} \quad (31)$$

The estimate $\widehat{\underline{b}}$ of $\underline{b}$ is obtained using Algorithm 3. As $\widehat{\mathbf{B}}$ is upper triangular, the estimation of $\underline{b}$ in $\widehat{r_b} = \widehat{\mathbf{B}}\underline{b}$ is fast performed by back substitution, instead of inverting a full matrix. The algorithm convergence was observed after about 20 iterations.

Algorithm 3 Iterative estimation of noise parameters

estimate the autocorrelation vector $\underline{r_b}$ by the empiric correlation coefficients $\widehat{r_b}$ obtained from the spectral observations $\{X(f)\}_{f\in\mathcal{F}_b}$;
initialize $\widehat{\underline{b}} \leftarrow (1, 0, \ldots, 0)^{\mathrm{T}}$;
for each iteration **do**
  update the estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$ from $\widehat{\underline{b}}$;
  re-estimate $\widehat{\underline{b}}$ by solving $\widehat{r_b} = \widehat{\mathbf{B}}\widehat{\underline{b}}$;
  normalize $\widehat{\underline{b}}$ by its first coefficient;
end for

REFERENCES

[1] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 1, pp. 24–33, 1977.

[2] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acous. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acous. Soc. Amer.*, vol. 43, no. 4, pp. 829–834, 1968.

[4] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method," *J. Acous. Soc. Amer.*, vol. 92, no. 3, pp. 1394–1402, 1992.

[5] B. Doval and X. Rodet, "Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, vol. 1, Apr. 1993, pp. 221–224.

[6] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Toulouse, France, May 2006, pp. 53–56.

[7] V. Emiya, B. David, and R. Badeau, "A parametric method for pitch estimation of piano tones," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Honolulu, HI, USA, Apr. 2007, pp. 249–252.

[8] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[9] C. Yeh, A. Robel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Philadelphia, PA, USA, Mar. 2005, pp. 225–228.

[10] H. Kameoka, T. Nishimoto, and S. Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[11] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, ser. Synthesis lectures on speech and audio processing, B. Juang, Ed. Morgan and Claypool Publishers, 2009.

[12] C. Raphael, "Automatic transcription of piano music," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 15–19.

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

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

[15] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Springer, 2006.

[16] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, vol. 8, pp. 1–9, 2007.

[17] K. Robinson and R. Patterson, "The duration required to identify the instrument, the octave, or the pitch chroma of a musical note," *Music Perception*, vol. 13, no. 1, pp. 1–15, 1995.

[18] L. I. Ortiz-Berenguer, F. J. Casajús-Quirós, and M. Torres-Guijarro, "Non-linear effects modeling for

polyphonic piano transcription," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, London, UK, Sep. 2003.

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

[20] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. Springer, 1998.

[21] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of inharmonic sounds in colored noise," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Bordeaux, France, Sep. 2007, pp. 93–98.

[22] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, 1991.

[23] R. Badeau and B. David, "Weighted maximum likelihood autoregressive and moving average spectrum modeling," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Las Vegas, NV, USA, Mar.-Apr. 2008, pp. 3761–3764.

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

[25] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Bordeaux, France, Sep. 2007, pp. 237–244.

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

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

[28] P. Walmsley, S. Godsill, and P. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters," in *Proc. IEEE Work. Appli. Sig. Proces. Audio and Acous. (WASPAA)*, New Paltz, NY, USA, Oct. 1999, pp. 119–122.

[29] V. Emiya, "Transcription automatique de la musique de piano," Ph.D. dissertation, École Nationale Supérieure des Télécommunications, France, 2008.

[30] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[31] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[32] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

Valentin Emiya was born in France in 1979. He graduated from Telecom Bretagne, Brest, France, in 2003 and received the M.Sc. degree in Acoustics, Signal Processing and Computer Science Applied to Music (ATIAM) at Ircam, France, in 2004. He received his Ph.D. degree in Signal Processing in 2008 at Telecom ParisTech, Paris, France. Since November 2008, he is a post-doctoral researcher with the METISS group at INRIA, Centre Inria Rennes - Bretagne Atlantique, Rennes, France.

His research interests focus on audio signal processing and include sound modeling and indexing, source separation, quality assessment and applications to music and speech.

Roland Badeau (M'02) was born in Marseille, France, on August 28, 1976. He received the State Engineering degree from the École Polytechnique, Palaiseau, France, in 1999, the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001, the M.Sc. degree in applied mathematics from the École Normale Supérieure (ENS), Cachan, France, in 2001, and the Ph.D degree from the ENST in 2005, in the field of signal processing. In 2001, he joined the Department of Signal and Image Processing, TELECOM ParisTech (ENST), as an Assistant Professor, where he became Associate Professor in 2005. His research interests include high resolution methods, adaptive subspace algorithms, audio signal processing, and music information retrieval.

Bertrand David (M'06) was born on March 12, 1967 in Paris, France. He received the M.Sc. degree from the University of Paris-Sud, in 1991, and the Agrégation, a competitive French examination for the recruitment of teachers, in the field of applied physics, from the École Normale Supérieure (ENS), Cachan. He received the Ph.D. degree from the University of Paris 6 in 1999, in the fields of musical acoustics and signal processing of musical signals.

He formerly taught in a graduate school in electrical engineering, computer science and communication. He also carried out industrial projects aiming at embarking a low complexity sound synthesizer. Since September 2001, he has worked as an Associate Professor with the Signal and Image Processing Department, GET-Télécom Paris (ENST). His research interests include parametric methods for the analysis/synthesis of musical and mechanical signals, spectral parametrization and factorization, music information retrieval, and musical acoustics.