BLIND HARMONIC ADAPTIVE DECOMPOSITION APPLIED TO SUPERVISED SOURCE SEPARATION

Benoit Fuentes, Roland Badeau, Gaël Richard

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
37-39, rue Dareau - 75014 Paris - France
benoit.fuentes@telecom-paristech.fr

ABSTRACT

In this paper, a new supervised source separation system is introduced. The Constant-Q Transform (CQT) of an audio signal is first analyzed through an algorithm called Blind Harmonic Adaptive Decomposition (BHAD). This algorithm provides an estimation of the polyphonic pitch content of the input signal, from which the user can select the notes to be extracted. The system then automatically separates the corresponding source from the audio mixture, by means of time-frequency masking of the CQT. The system has been evaluated both in a task of multipitch estimation in order to measure the quality of the decomposition, and in a task of user-guided melody extraction to assess the quality of the separation. The very promising results obtained highlight the reliability of the proposed model.

Index Terms— Audio Source Separation, Harmonic Decomposition, PLCA, NTF

1. INTRODUCTION

Spectrogram decomposition techniques into meaningful basic elements such as Nonnegative Matrix Factorization are now widely used to perform monaural source separation of audio mixtures [1]. However, performing this task in a completely blind way remains challenging, basically due to the difficulty of clustering the basic elements that belong to the same source. To overcome this problem, one solution is to inform the separation, e.g. whether with the aligned score of the audio [2] or with user-specifiable constraints over the present sources in the mixture [3].

In this paper, another approach is followed, similar to [4], where no side information is needed (the decomposition is blindly performed) but where the user can cluster the basic elements of the sources to be separated. In fact, the proposed system allows the user choosing the notes to be extracted via an intuitive Graphical User Interface (GUI) showing an estimation of the polyphonic pitch content of the input signal. This work relies on [5], where it was proven that efficient translation-invariant models for music analysis on the CQT of an audio signal can be directly applied to source separation. Thus, the article is organized as follows: first, the statistical framework that is employed to model the noise and the polyphonic part of a CQT is described in section 2. We present the algorithm for estimating the model parameters in section 3. In section 4, a new sparseness prior is introduced in order to constrain the parameters to converge towards a meaningful solution. The application to source separation is exposed in section 5, where the GUI is described and the separation process explained. Finally, we present two different evaluations of the method in section 6 and conclude in section 7.

The research leading to this paper was partly supported by the Quaero Programme, funded by OSEO, French State agency for innovation.

2. PROBABILISTIC MODEL FOR THE INPUT CQT

The framework on which the presented model relies is the Probabilistic Latent Component Analysis (PLCA) following the example of [6]. It is a probabilistic tool for non-negative data analysis which offers a convenient way of designing spectrogram models and introducing priors on the corresponding parameters. Let us consider the absolute value of the CQT X_{ft} of an audio signal x, denoted $V_{ft} = |X_{ft}|$. In PLCA, it is modeled as the histogram of N independent random variables $(f_n, t_n) \in \mathbb{Z} \times [\![1; T]\!]$, which represent time-frequency bins, distributed according to $P(f,t)$ (we suppose that $V_{ft} = 0$ for $f \notin [\![1, F]\!]$). $P(f,t)$ can then be parameterized according to the desired decomposition of V_{ft} and the parameters can be found by means of the Expectation-Maximization (EM) algorithm. In the inherent model of the BHAD algorithm (that we call the BHAD model), a first latent variable c is introduced in order to decompose V_{ft} as a sum of a polyphonic harmonic signal (in this case, $c = h$) and a noise signal ($c = n$) (the notations $P_h(.)$ and $P_n(.)$ are used for $P(.|c = h)$ and $P(.|c = n)$):

$$P(f,t) = P(c = h)P_h(f,t) + P(c = n)P_n(f,t), \quad (1)$$

where $P_h(f,t)_{(f,t)\in\mathbb{Z}\times[\![1,T]\!]}$ and $P_n(f,t)_{(f,t)\in\mathbb{Z}\times[\![1,T]\!]}$ respectively represent the normalized CQTs of the polyphonic and the noise signals. The models for each component are presented in the following sections.

2.1. The notes model

At a given time t, the spectrum of the polyphonic component, represented by $P_h(f,t)$, is decomposed as a weighted sum of different harmonic notes, each one having its own fundamental frequency (on a discrete logarithmic scale $\mathrm{pitch}(i)_{i\in[\![0,I-1]\!]}$) and spectral envelope. Since the number of notes is unknown, we consider all possible fundamental frequencies, with possibly zero weights:

$$P_h(f,t) = \sum_i P_h(i,t)P_h(f|i,t), \tag{2}$$

where $i \in [\![0, I-1]\!]$ is a new latent variable representing the note of fundamental frequency $\mathrm{pitch}(i)$. $P_h(i,t)$ and $P_h(f|i,t)$ respectively represent the energy and the normalized harmonic spectrum of note i at time t. $P_h(i,t)$ can thus be seen as time-frequency note activations.

In order to account for the harmonic nature of $P_h(f|i,t)$, as well as its specific spectral envelope, the same principle as in [7] is adopted. The spectrum of a harmonic note is decomposed as a weighted sum of Z fixed narrow-band harmonic spectral kernels, denoted $P_h(f|z,i)$, sharing the same fundamental frequency $\mathrm{pitch}(i)$ and having their energy concentrated at the z^{th} harmonic:

$$P_h(f|i,t) = \sum_z P_h(z|i,t)P_h(f|z,i). \tag{3}$$

The applied weights, included in $P_h(z|i,t)$, define the spectral envelope of the current note. Working with the CQT has a main advantage, since for a harmonic note, a pitch modulation can be interpreted as a frequency shifting of the partials. Therefore, for given i and z, $P_h(f|z,i)$ can be deduced from a single template $P_h(\mu|z)_{\mu\in[\![1,F]\!]}$ as follows:

$$P_h(f|z,i) = P_h(f-i|z). \tag{4}$$

Here, $P_h(\mu|z)$ is also a narrow-band harmonic kernel, having its energy concentrated on the z^{th} harmonic. Its fundamental frequency is $\mathrm{pitch}(0)$. Its precise definition can be seen in the function *make_Kernel.m* of the online Matlab code [8]. Contrary to [9], the kernels have been designed in order to have energy only for the frequency bins corresponding to harmonics. The spectral spreading of the partials of a note is therefore not taken into account by the kernels but by the note activations, various joint values of i being necessary to explain a single note. Doing so allows insuring that the model can fit any spectral spreading of the partials (for instance, a continuous variation of pitch induces a larger spreading at a given time).

Finally, the whole polyphonic component model can be written as:

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

One can notice that we end up with a convolutive model, meaning that variable f is defined as the sum of two random variables μ and i. Fig. 1 illustrates this model.

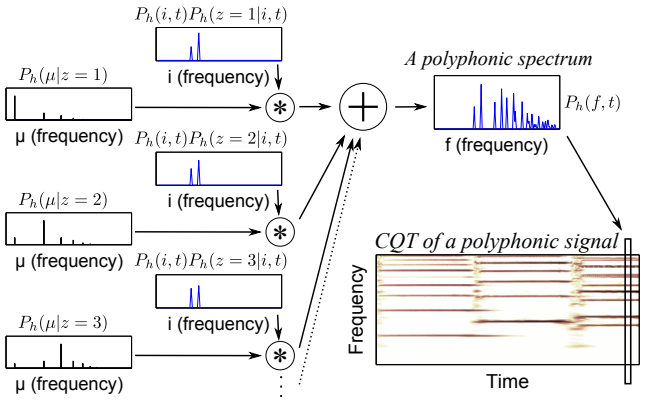

Fig. 1. Polyphonic component of the BHAD model. At time t_0, the vector $P_h(i,t_0)$ should have as many peaks as there are active notes in the signal.

2.2. The noise model

Similarly to [7], the CQT of the noise signal is modeled as the convolution of a fixed smooth narrow-band window $P_n(\mu)_{\mu\in[\![1,F]\!]}$, and a noise time-frequency distribution $P_n(i,t)_{(i,t)\in[\![0,I-1]\!]\times[\![1,T]\!]}$:

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

3. PARAMETERS ESTIMATION: EM ALGORITHM

In [7], it is explained how to derive the EM algorithm. This algorithm defines update rules for the parameters so that the log-likelihood L of the observations increases at every iteration (it can be proven that $L = \sum_{f,t} V_{ft} \ln\left(P(f,t)\right)$).

First, in the "expectation step", the posterior distribution of latent variables i, z and c is computed by applying the Bayes' theorem:

$$P(i,z,c=h|f,t) = \frac{P(c=h)P_h(i,t)P_h(z|i,t)P_h(f-i|z)}{P(f,t)}, \tag{7}$$

$$P(i,c=n|f,t) = \frac{P(c=n)P_n(i,t)P_n(f-i)}{P(f,t)}. \tag{8}$$

Equations (1), (5) and (6) define $P(f,t)$.

Then, in the "maximization step", the log-likelihood of observed and latent variables Q_Λ is maximized, leading to the following updates rules:

$$P(c=h) \propto \sum_{f,t,z,i} V_{ft}P(i,z,c=h|f,t), \tag{9}$$

$$P_h(i,t) \propto \sum_{f,z} V_{ft}P(i,z,c=h|f,t), \tag{10}$$

$$P_h(z|i,t) \propto \sum_f V_{ft} P(i,z,c=h|f,t), \quad (11)$$

$$P(c=n) \propto \sum_{i,f,t} V_{ft} P(i,c=n|f,t), \quad (12)$$

$$P_n(i,t) \propto \sum_f V_{ft} P(i,c=n|f,t). \quad (13)$$

After initialization of the parameters, the EM algorithm consists in iterating equations (7) and (8), the different update rules (equations (9) to (13)) and finally the normalization of all parameters so that the probabilities sum to one.

4. SPARSENESS PRIOR

In practice, running the presented algorithm without any additional prior does not give relevant estimations of the parameters. Indeed, for one note of pitch f_0 actually present in the input signal, all notes i whose fundamental frequency is a multiple or a submultiple of f_0 will be activated. Thus, even if the log-likelihood of the data after convergence is high, the decomposition might not be informative enough. In order to overcome this flaw, we add a sparseness prior to the note activations $P_h(i,t)$, assuming it is more likely to explain the same set of data using a fewer number of active notes.

If $\boldsymbol{\theta}$ is the $I \times T$ matrix of coefficients $\theta_{it} = P_h(i,t)$, the prior we put forward is defined as follows:

$$Pr(\boldsymbol{\theta}) \propto \exp\left(-2\beta\sqrt{IT}\,\|\boldsymbol{\theta}\|_{1/2}\right). \quad (14)$$

where $\|\boldsymbol{\theta}\|_{1/2} = \sum_{i,t}\sqrt{\theta_{it}}$. β is a hyperparameter defining the strength of the prior and the coefficient $\sqrt{IT}$ is such that the strength is independent of the size of the data. In Appendix A, it is proven that if $\beta^2 < \sum_{i,t} w_{it}^2/(IT)$, where $w_{it} = \sum_{f,z} V_{ft} P(i,z,c=h|f,t)$, then the update rule (10) followed by its normalization are replaced by:

$$P_h(i,t) = \frac{2w_{it}^2}{IT\beta^2 + 2\rho w_{it} + \beta\sqrt{IT}\sqrt{IT\beta^2 + 4\rho w_{i,j}}}, \quad (15)$$

ρ being the unique positive number such that $P_h(i,t)$ sums to one. This number can be found with any root finder algorithm (we used the *fzero* Matlab function). In practice β is set to a sufficiently low value so that β^2 is always inferior to $\sum_{i,t} w_{it}^2/(IT)$. The effect of using the sparseness prior is illustrated in Fig. 2.

5. APPLICATION TO SUPERVISED SOURCE SEPARATION

In this section, it is explained how the BHAD model can be used to perform supervised source separation. The main idea is to provide a note selection tool where the user chooses via a GUI which notes actually present in the input file are to be separated from the rest of the audio.

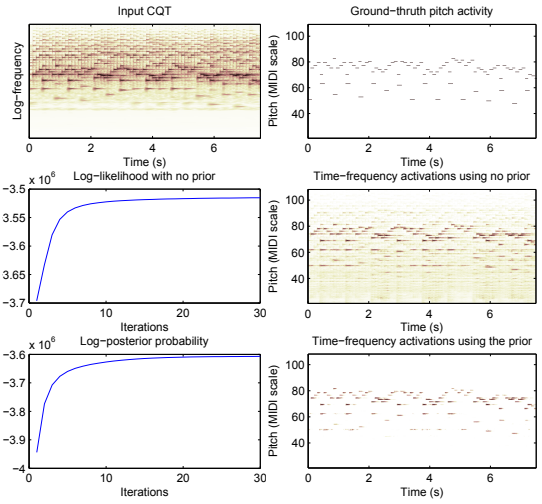

Fig. 2. Illustration of the use of the sparseness prior. The input signal corresponds to an excerpt from Bach's *Prelude and Fugue in D major BWV 850*. The growth of the criterion over the iterations of the EM algorithm has also been plotted.

5.1. GUI and notes selection

The BHAD model offers a relevant mid-level representation of audio, since note activations $P_h(i,t)$ indicate the active notes with respect to time, like in a "piano-roll" representation. Using Matlab, a GUI has been developed (available at [8]), as shown in Fig. 3, where the user can highlight the notes to be extracted. The GUI consists of the following elements: (1) the representation of note activations $P_h(i,t)$, on which the user can select notes or edit the data (erase and draw functions) if he notices that the BHAD algorithm gave wrong estimations; (2) the control panel, where the user can set hyperparameters for the BHAD algorithm (such as the sparseness strength β), run the algorithm, separate the selected notes, reinitialize all parameters and change the contrast of the representation of the activations; (3) the toolbar, composed of basic tools in order to load a new wave file or a previous work, to save the current work, to explore the image, to select or unselect notes, to listen to a note whose pitch corresponds to the same pitch than the selected note, to edit the note activations and finally to listen to the signal.

5.2. Source model and time-frequency masking

The user, by highlighting the notes he wants to extract with the "select note" tool, defines a binary mask $B(i,t)$ on the note activations, equal to 1 if a time-frequency bin is selected and 0 otherwise. $B(i,t)$ can be used to perform the separation by means of time-frequency masking on the input CQT. Two masks, M_1 and M_2, which respectively correspond to source 1

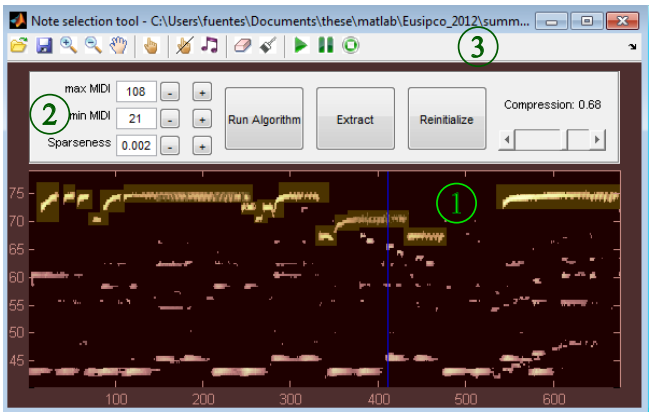

Fig. 3. GUI of the note selection tool. The input file is an excerpt from the jazz standard *Summertime* composed by George Gershwin. The highlighted areas correspond to notes selected by the user.

(the selected notes) and source 2 (the residual), are defined as:

$$M_1(f,t) = \frac{P(c=h)\sum_{i,z} B_1(i,t)P_h(z|i,t)P_h(f-i|z)}{P(f,t)}, \tag{16}$$

$$M_2(f,t) = \frac{1}{P(f,t)}\Big(P(c=n)P_n(f,t)+ \tag{17}$$
$$P(c=h)\sum_{i,z} B_2(i,t)P_h(z|i,t)P_h(f-i|z)\Big),$$

where $B_1(i,t)$ and $B_2(i,t)$ respectively denote $B(i,t)P_h(i,t)$ and $(1-B(i,t))\,P_h(i,t)$. It can be noticed that for all (f,t), $M_1(f,t)+M_2(f,t)=1$. The estimated temporal signals of the two sources, $\hat{x}_1$ and $\hat{x}_2$, are then given by applying the masks on the input CQT X_{ft} and calculating the invert CQT[1]:

$$\hat{x}_1 = \mathrm{CQT}^{-1}\left(M_1(f,t)X_{ft}\right), \tag{18}$$
$$\hat{x}_2 = \mathrm{CQT}^{-1}\left(M_2(f,t)X_{ft}\right). \tag{19}$$

6. EVALUATION

Two different evaluations have been made in order to measure the quality of the presented system. The aim of the first evaluation is to appreciate the relevance of the BHAD algorithm itself, in order to ensure that it gives accurate estimation of note activations. Thus, it has been evaluated in a task of multipitch estimation.

First, the CQT of a temporal signal is calculated from $f = 27.5$Hz to $f = 7040$Hz with 3 frequency bins/semitones and with a time step of 10 ms. After convergence of the BHAD algorithm, the pitches are inferred for each time frame from the note activations: at a given time t_0, the note i_0 is considered to be active if $P_h(i,t_0)$ presents a local maximum in i_0 and if $P_h(i_0,t_0)_{\mathrm{dB}} > \max_{i,t} P_h(i,t)_{\mathrm{dB}} - A_{min}$. Finally the corresponding fundamental frequency $\mathrm{pitch}(i_0)$ is rounded to the closest MIDI pitch. In order to evaluate the role of the sparseness prior, two versions of the algorithm have been tested. BHAD-np will refer to the system using no prior ($\beta = 0,\ A_{min} = 25$dB), and BHAD to the system with the prior ($\beta = 0.0018,\ A_{min} = 30$dB). The values of β and A_{min} have been set according to the results obtained during a training process on a development database. The test database used for evaluation is a subset of the QUAERO database[2] (6 audio files of various genre, from reggae to rock) and the MIREX 2007 multi-F0 development dataset[3]. The algorithm [9] has also been evaluated. Four classical measures, Precision, Recall, F-measure and Accuracy, described in [9] and [10], are reported in Tab. 1. It can be seen that the addition of the sparseness prior significantly improves the precision of the results, despite a lower score in terms of recall. In any case, the BHAD algorithm outperforms the reference algorithm for every measure.

Algorithm	Precision	Recall	F-measure	Accuracy
[9]	29.3	53.2	35.8	81.7
BHAD-np	30.0	**64.2**	31.2	76.6
BHAD	**47.0**	54.5	**47.2**	**85.3**

Table 1. Results from QUAERO 2011 framewise multipitch evaluation task.

The second evaluation is dedicated to the user-guided source separation, where the proposed system was compared to [4] in a task of main melody (vocal source) extraction. The database consists of five 15s excerpts from the QUAERO source separation corpus. For each file and each system, the main melody pitch line has been localized, selected by means of the selection tool provided in both GUIs and finally extracted. The quality of the estimated melody source is then quantified through the BSSEval toolbox [11], which gives the following measures: the Signal to Distortion Ratio (SDR), the Signal to Interference Ratio (SIR) and the Signal to Artifact Ratio (SAR). According to the results reported in Tab. 2, it seems that our system is slightly less efficient in a task of melody extraction. However, it can be noticed that it is more generic since it allows separating any polyphonic source whereas in [4], only monophonic sources can be extracted.

[1]The invert CQT we used in our algorithm is freely available at http://www.tsi.telecom-paristech.fr/aao/en/2011/06/06/inversible-cqt

[2]The QUAERO (http://www.quaero.org) database will be soon available at http://www.tsi.telecom-paristech.fr/aao/en/software-and-database/

[3]http://music-ir.org/mirexwiki

Method	SDR	SIR	SAR
Proposed	4.0	**16.6**	4.5
[4]	**5.2**	16.2	**6.0**

Table 2. Average SDR, SIR and SAR of the melody estimates using two systems.

7. CONCLUSION

In this study, we propose a new algorithm which accurately decomposes the CQT of an audio signal into a meaningful mid-level representation. Each time frame of the CQT is decomposed as a sum of harmonic notes, each note being modeled by means of fixed narrow-band harmonic templates. The presence of colored noise is also considered, and a new sparseness prior has been introduced for note activations. The BHAD algorithm has been evaluated in a task of multiple pitch estimation, and outperformed another state-of-the art algorithm. Finally, it has been proven that the BHAD model could be used for source separation, by offering a GUI where the user can select the notes he wants to extract. In future work, the authors plan to include a noise model for the notes, since for now, only noiseless harmonic instruments can be correctly being separated. Another outlook would be to automatically cluster the notes according to their timbre in order to isolate instruments in an unsupervised way.

A. EM UPDATE RULES WITH SPARSE PRIOR

During the "maximization step" of the EM algorithm, one wants to maximize the joint log-probability Q_Λ of all variables (Λ denotes the set of all parameters). Using the same notation as in section 4, it can be proven that

$$Q_\Lambda = Q_{\Lambda'} + \sum_{i,t} w_{it} \ln(\theta_{it}) \tag{20}$$

where $Q_{\Lambda'}$ depends on all parameters other than $\theta_{it} = P_h(i,t)$. With the addition of the sparseness prior, the maximization step is now replaced by a maximization a posteriori step. It does not change anything for the other parameters, but the new update rule for θ_{it} is obtained by maximizing $Q_\Lambda + \ln(Pr(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$. It amounts to maximizing on $\Omega =]0,1]^I \times]0,1]^T$ the following functional under the constraint $\varphi(\boldsymbol{\theta}) = 1 - \sum_{i,t} \theta_{it} = 0$:

$$\begin{aligned} \mathcal{S} \,:\, \Omega &\longrightarrow \mathbb{R} \\ \boldsymbol{\theta} &\longmapsto \sum_{i,t} w_{it} \ln(\theta_{it}) - 2\beta\sqrt{IT} \sum_{i,t} \sqrt{\theta_{it}}. \end{aligned} \tag{21}$$

We know that the maximum exists on Ω since S is continuous and upper bounded by 0 and its argument $\hat{\boldsymbol{\theta}}$ verifies the first order necessary conditions, proper to local maxima (Lagrange theorem): since $\mathcal{S}$ and φ are both differentiable, there exists a unique $\rho \in \mathbb{R}$ such that:

$$\nabla L_\rho(\hat{\boldsymbol{\theta}}) = 0 \tag{22}$$

where L_ρ is the Lagrangian defined as:

$$\begin{aligned} L_\rho \,:\, \Omega &\longrightarrow \mathbb{R} \\ \boldsymbol{\theta} &\longmapsto \mathcal{S}(\boldsymbol{\theta}) + \rho\,\varphi(\boldsymbol{\theta}). \end{aligned} \tag{23}$$

Equation (22) leads to:

$$\forall (i,t),\ \frac{w_{it}}{\hat{\theta}_{it}} - \frac{\beta\sqrt{IT}}{\sqrt{\hat{\theta}_{it}}} - \rho = 0. \tag{24}$$

By studying the three cases $\max_{i,t}\left(-\frac{\beta^2 IT}{4w_{it}}\right) < \rho < 0$, $\rho = 0$ and $\rho > 0$, it appears that $\sum_{i,t} \frac{w_{it}^2}{\beta^2 IT} > 1$ if and only if $\rho > 0$. In that event,

$$\forall (i,t),\ \hat{\theta}_{it} = \frac{2w_{it}^2}{IT\beta^2 + 2\rho w_{it} + \beta\sqrt{IT}\sqrt{\beta^2 IT + 4\rho w_{it}}}, \tag{25}$$

ρ being the unique value for which $\sum_{i,t} \hat{\theta}_{it} = 1$. This finally proves eq.(15).

8. REFERENCES

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

[2] S. Ewert and M. Müller, "Score informed source separation," in *Multimodal Music Processing*, Masataka Goto Meinard Müller and Markus Schedl, Eds., Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2012.

[3] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, May 2012.

[4] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. of LVA/ICA*, Tel-Aviv, Israel, March 2012.

[5] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard, "Probabilistic model for main melody extraction using constant-Q transform," in *Proc. of ICASSP*, Kyoto, Japan, March 2012.

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

[7] B. Fuentes, R. Badeau, and G. Richard, "Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA," in *Proc. of ICASSP*, Prague, Czech Republic, May 2011, pp. 401–404.

[8] "Companion website," http://www.tsi.telecom-paristech.fr/aao/?p=756.

[9] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio Speech and Language Processing*, 2010.

[10] S. Dixon, "On the computer recognition of solo piano music," in *Proc. of Australasian Computer Music Conference*, July 2000, pp. 31–37.

[11] E. Vincent, C. Févotte, and R. Gribonval, "Performance measurement in blind audio source separation," *Audio, Speech and Language Processing, IEEE Trans. on*, vol. 14, no. 4, pp. 1462–1469, 2006.