CONTROLLING THE CONVERGENCE RATE TO HELP PARAMETER ESTIMATION IN A PLCA-BASED MODEL

Benoit Fuentes, Roland Badeau, Gaël Richard

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

ABSTRACT

Probabilistic Latent Component Analysis (PLCA) is a tool similar to Non-negative Matrix Factorization (NMF), which is used to model non-negative data such as non-negative time-frequency representations of audio. In this paper, we put forward a trick to help the corresponding parameter estimation algorithm to converge toward more meaningful solutions, based on the new concept of brakes. The idea is to control the convergence rate of the parameters of a PLCA-based model within the estimation algorithm: the parameters which are known to be properly initialized are braked in order to stay close to their initial values, whereas the other ones keep a regular convergence rate. This is an effective way to better account for a relevant initialization. In this paper, these brakes are implemented in the framework of PLCA, and they are tested in an application of multipitch estimation. Results show that the use of brakes can significantly influence the decomposition and thus the performance, making them a powerful tool to boost any kind of PLCA-based algorithm.

***Index Terms*—** PLCA, NMF, EM algorithm, multipitch estimation.

1. INTRODUCTION

Factorizations of non-negative time-frequency representations (TFR$^+$) are used in many audio applications such as multipitch estimation, automatic transcription and source separation. They require to put forward models of TFR$^+$ and to find algorithms to estimate the model parameters given an observation. To this aim, many mathematical frameworks can be used, either deterministic [1, 2] or probabilistic [3, 4, 5]. Those methods, though quite efficient, may also suffer from several flaws. Indeed, depending on the model and algorithm considered, the following problems are often faced:

- *Unidentifiability* of the model: several sets of parameters can explain the same observation.
- *Local optima*: the algorithm may stay stuck in a sub-optimal local optimum.
- *Relevancy* of the solution: the optimal solution is not necessarily the most meaningful one.

In order to overcome these problems, an often used solution is the addition of penalty terms or priors on the parameters. For instance, many studies have been conducted in order to enforce sparsity or temporal continuity of a subset of parameters for a given model [6, 7].

In this paper, we study a new idea, which, as well as the addition of prior or penalty terms, helps parameter estimation algorithms to converge towards meaningful solutions. The idea is to slow down the rate of convergence of the parameters that are well initialized, through the use of a brake[1]. By doing so, it is ensured that after convergence, the estimates of the braked parameters stay close to their initial values. The other parameters — the value of which we have no prior knowledge on — keep a regular convergence rate. In the same way a sled turns right if only the right brake is used, this approach is a simple way to affect the direction that the algorithm takes and make it converge to a more relevant local optimum, by better accounting for a proper initialization.

In order to explore this idea, we choose to focus on the mathematical framework of PLCA with its basic model [4], that we recall in section 2. In section 3, the brakes are introduced in order to independently control the convergence rate of different sets of parameters. Experimental studies are then conducted in section 4, through the problem of multipitch estimation. Finally, conclusions are drawn in section 5.

2. PLCA

PLCA [4] is a method for analyzing non-negative data: here the non-negative coefficients that compose a TFR$^+$ $\boldsymbol{V}$ of an input audio signal. The observations $\boldsymbol{V}$ of coefficients V_{ft}, $t \in [\![1, T]\!]$ and $f \in [\![1, F]\!]$ being respectively time and frequency indexes, are modeled as the histogram of the sampling of J independent random variables (r.v.) (f, t). These variables are drawn according to the probability distribution $P_\Lambda(f, t)$, and the way this distribution is modeled induces the

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

[1] The word "brake" is a new term we put forward in this context, defined in section 3.

wanted decomposition. The brakes, introduced in next section, could be used for any observation model, but for the sake of simplicity, only the basic model [4] is considered here. In this model, a latent variable $n \in [\![1, N]\!]$ is introduced (it can for instance represent a MIDI note), f and t are considered independent conditionally to n and $P_\Lambda(f,t)$ is modeled as:

$$P_\Lambda(f,t) = \sum_n P(n,t)P(f|n). \quad (1)$$

The set of parameters Λ is defined as $\{P(n,t), P(f|n)\}_{n,t,f}$. In this paper, we suppose that $P(f|n)$ represents the basis spectrum of note n (also called atom) and $P(n,t)$ its time activation.

Given an observation, the parameters can be estimated by means of the Expectation-Maximization (EM) algorithm [8]. It provides update rules for the parameters so that the likelihood of the observations does not decrease at each iteration j:

$$P(n,t)^{j+1} \propto P(n,t)^j \sum_f \frac{V_{ft}}{P_{\Lambda^j}(f,t)} P(f|n)^j \quad (2)$$

$$P(f|n)^{j+1} \propto P(f|n)^j \sum_t \frac{V_{ft}}{P_{\Lambda^j}(f,t)} P(n,t)^j \quad (3)$$

where $\propto$ means "proportional to" (*i.e.* the parameters must be normalized after their updates, so that the probabilities sum to 1) and where $P_\Lambda(f,t)$ is defined in equation (1). From now on and for the sake of clarity, we deliberately omit to specify that a given distribution depends on the parameters Λ and that the value of a parameter depends on the iteration j of the algorithm. For instance, $P_{\Lambda^j}(f,t)$ is thus denoted $P(f,t)$.

3. TAKING CONTROL OF THE CONVERGENCE RATE

Many works have been done in order to influence the convergence rate of NMF-based algorithms (e.g. [9]), but their goal is to study or improve the overall convergence speed. In this section, we suggest to independently control the convergence rate of different sets of parameters ($\{P(n,t)\}_{n,t}$ and $\{P(f|n)\}_{f,n}$ for instance in the case of basic PLCA model) in order to influence the result of an optimization algorithm, without looking at the resulting global speed convergence. A simple idea to do so is to compute a weighted mean between the optimal solution of the M step and the value of the parameters at the previous iteration. Equations (2) and (3) would then become:

$$P(n,t) \propto P(n,t) \left[\sum_f \frac{V_{ft}}{P(f,t)} P(f|n) + \beta^1_{\text{brake}} \right], \quad (4)$$

$$P(f|n) \propto P(f|n) \left[\sum_t \frac{V_{ft}}{P(f,t)} P(n,t) + \beta^2_{\text{brake}} \right], \quad (5)$$

where β^1_{brake} and β^2_{brake} are two positive coefficients, that we call *brakes*. They indeed act as a brake on the convergence, since the larger they are, the more the value of the parameters at a given iteration is close to the previous iteration. If the two brakes are set with different values (for instance $\beta^1_{\text{brake}} = 0$ and $\beta^2_{\text{brake}} > 0$), they will act as a "steering wheel", and then influence the direction in which the algorithm goes. The parameters may then converge towards a different local minimum than if no brakes were use, as shown in next section. For instance, one use that can be put forward is to brake the parameters which are known to be well initialized. In this way, it is more likely that, after convergence, they will not be far from their initialization.

There is a formal way to introduce the brakes in the framework of PLCA, so that equations (4) and (5) are directly derived from the EM algorithm. The reader is referred to appendix A for further explanations.

4. EXPERIMENTAL STUDY AND APPLICATION

In this section, we study how the use of brakes can influence the decomposition algorithm. To do so, we chose to address the problem of multipitch estimation. The purpose of this section is not to provide new state of the art algorithms but to show that the use of brakes can improve the performance of existing methods, whether they are naive or sophisticated. By doing so, we hope that other researchers will be able to use brakes in order to improve the parameter estimation algorithm of their own models, whatever the problem considered.

From now on, the TFR$^+$ used is the absolute value of the constant-Q transform (CQT) with 3 bins/semitone, for frequencies from 27.5 to 7040 Hz and with a time step of 10 ms. Besides, only piano signals are considered since for the piano, the assumption that each note can be modeled by a single basis spectrum is a reasonable assumption. In all the following multipitch estimation applications, the dataset used is a subset of the MAPS database [10] composed of nine 20s excerpts. It is denoted as DB_{eval}. For each of the following algorithms, in order to set the values of the brakes when they are non-zero, another subset of MAPS of the same size is used as a training dataset.

4.1. Blind PLCA

The model of the first conducted experiment is similar to the one put forward in [11]. It concerns blind PLCA, *i.e.* when no knowledge is provided on the nature of the atoms $P(f|n)$ or the activations $P(n,t)$. To do so, $P(n,t)$ is initialized with a uniform distribution and $P(f|n)$ is initialized with some random distribution. Since the total number of different notes in an input signal is also unknown, we consider that all MIDI notes might be present and the number of atoms used to model them is thus set to 88 (the total number of keys on a piano). Moreover, 4 additional atoms are reserved to

model the possible presence of noise, which makes a total of $N = 92$ atoms. Fig. 1 illustrates how the use of a brake on $P(f|n)$ can influence the decomposition. In fact, in a given musical excerpt, it is unlikely to have all the 88 MIDI notes (especially if the signal is short), and N is then overestimated. If we let the algorithm converge without using brakes, all the atoms available will be used to model the input data, with for instance several atoms used to model a single note. If one wants to model each note by a single atom, it is possible to slow down the convergence rate of the basis spectra. In this way, the algorithm is informed that the solution for $P(f|n)$ must be close to initialization, i.e. that the lowest possible number of atoms must correspond to note spectra (harmonic spectra), the others keeping their initial shape.

It is interesting to note that temporal activations become then much sparser, as well as the overall energy of the atoms $P(n) = \sum_t P(n,t)$: many atoms are never activated and the overestimation of N is no longer a problem. However, it can be noticed that with the use of brakes, the convergence of the log-likelihood is slower.

In order to quantify the benefits of using the brake on $P(f|n)$ in the framework of blind PLCA, a multipitch evaluation is performed on DB_{eval}. After convergence of the algorithm, at time t_0, a note n_0 is considered to be active if $P(n_0, t_0)_{\text{dB}} > \max_{n,t} P(n,t)_{\text{dB}} - A_{\text{min}}$ where A_{min} is a detection threshold. The pitch of each atom $P(f|n)$ is estimated using a simple spectral sum. The quality of the resulting note activation estimation can then be quantified through the three classical measures of Recall, Precision and F-measure [10]. In Fig. 2, the average F-measure w.r.t. A_{min} is illustrated, whether the brake is used or not. It can be seen that the brake can significantly improve the results, and at the same time can reduce the necessity of a fine tuning of the A_{min} threshold.

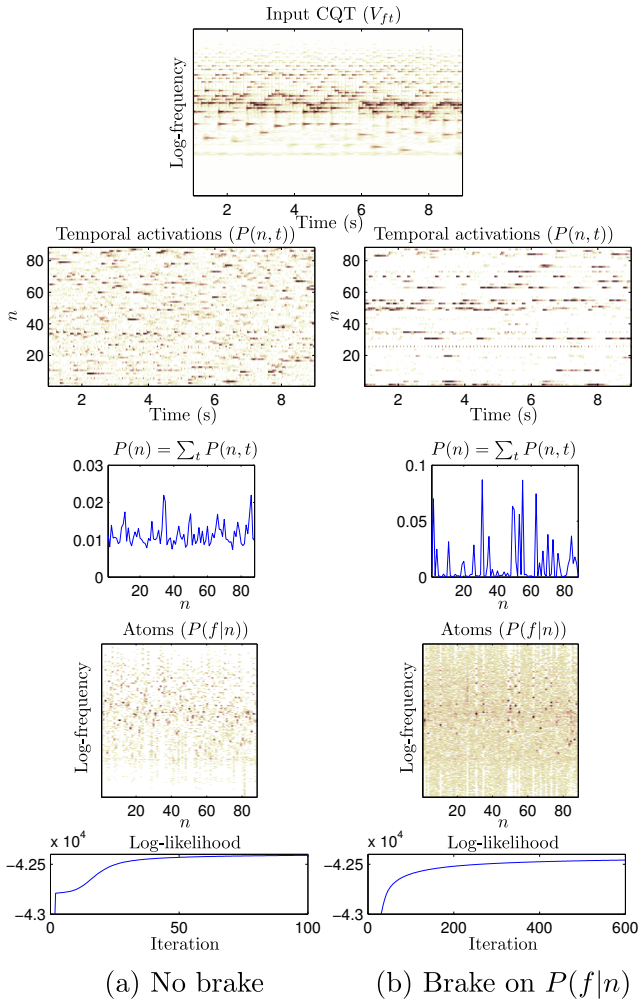

Fig. 1. Blind PLCA: illustration of the use of a brake on basis spectra ($\beta^1_{\text{brake}} = 0$ and $\beta^2_{\text{brake}} = 250$). The input signal is an 8s excerpt from the BWV 850 Bach's Prelude. The same random initialization of $P(f|n)$ has been used in the two experiments.

4.2. Harmonic PLCA

The second experiment is very similar to the previous one, with the same number of atoms used ($N = 92$). The only difference is that we add some knowledge on the nature of the signals to be analyzed by initializing the 88 first atoms as harmonic spectra. The coefficients of $P(f|n)$ are thus set to a very low value ϵ for frequencies between theoretical harmonics of note $n \in [\![1, 88]\!]$. Moreover, its spectral envelope is decreasing w.r.t. frequency, as usually observed in note spectra of acoustic instruments. This model and its initialization is thus similar to what is proposed in [12]. The four atoms used to model noise and activations are initialized in the same way as in the previous experiment.

Here again, a brake can be applied on the basis spectra since we can consider that they are properly initialized. In practice, we can notice that this induces sparser activations after convergence of the algorithm, as well as for blind PLCA. Here again, the use of brakes is evaluated through the same application of multipitch estimation on DB_{eval}. In this case, it is no longer necessary to estimate the pitch of each harmonic atom, since it has been established during initialization. Results are shown in Fig. 3, and the use of brakes again appears to be beneficial in terms of F-measure.

4.3. BHAD

Until now, only the basic PLCA model [4] has been considered. Though, as shown in this section, it is also possible to apply the principle of brakes to any kind of PLCA-based model, by using the same approach as in section 3. Here, we want to apply it to the Blind Harmonic Adaptive Decomposition (BHAD) [13]. It is a more sophisticated model which allows modeling harmonic notes having time variations of pitch and spectral envelope. The evaluation conducted in [13] has also shown that BHAD was at the state of the art for multipitch estimation. In

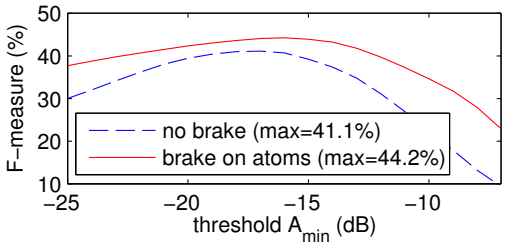

Fig. 2. Study of the influence of a brake on basis spectra in the framework of blind PLCA: average F-measure w.r.t. A_{min} for two multipitch estimation algorithms. When the brake is used, $\beta^1_{\text{brake}} = 0$ and $\beta^2_{\text{brake}} = 250$.

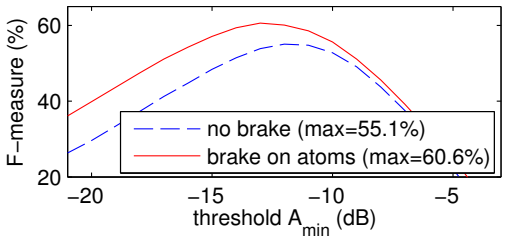

Fig. 3. Study of the influence of a brake on basis spectra in the framework of harmonic PLCA: average F-measure w.r.t. A_{min} for two multipitch estimation algorithms. When the brake is used, $\beta^1_{\text{brake}} = 0$ and $\beta^2_{\text{brake}} = 250$.

this model, each column of an input CQT is first decomposed into a smooth spectrum (noise) and a polyphonic harmonic spectrum. This last component is modeled as a weighted sum of various harmonic spectra, each one having its own pitch (denoted by a hidden variable i) and a time-dependent spectral envelope. In order to consider any number of active notes at a given time, all possible pitches are considered, with possible zero weights. The spectral envelope of each harmonic spectrum of pitch i at time t is encoded via a set of coefficients, which we call envelope coefficients and which are denoted $P_h(z|i,t)$ in [13].

The BHAD model is a very expressive model, and it is thus necessary to constrain the parameter estimation algorithm so that it gives relevant solutions. In [13], a sparse prior is added to the weights of the harmonic spectra and it is shown that the performance is improved when BHAD is applied to multipitch estimation. Here, in order to constrain the decomposition, a brake on the envelope coefficients is added: we force the spectral envelopes of the harmonic spectra to be close to their initialization (decreasing in frequency). Results of multipitch estimation on DB_{eval} are shown in Fig. 4. They show that the use of brakes in the BHAD model can also significantly improve the performance.

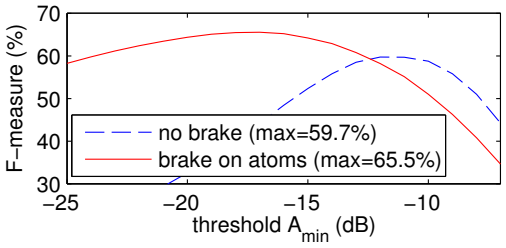

Fig. 4. Study of the influence of a brake on envelope coefficients in the framework of BHAD: average F-measure w.r.t. A_{min} for two multipitch estimation algorithms. When the brake is used, $\beta_{\text{brake}} = 10$.

5. CONCLUSIONS

In this paper, we have introduced a new way of helping a parameter estimation algorithm converge towards a more meaningful solution in the framework of PLCA-based models. This is done by slowing down the convergence rate of parameters that are properly initialized — the other ones keeping a regular convergence rate —, in order to better account for the initialization. In other words, it is a simple way to introduce prior knowledge on the signals to be analyzed. A major advantage of the use of brakes is that they are easily implemented and that they do not increase the complexity, even if they can also slow down the overall convergence of the algorithm. Three multipitch estimation algorithms have been tested with and without brakes, and it appears that brakes have significantly improved the performance in all cases. In future work, we plan to theorize the use of brakes for other kind of mathematical frameworks ([1, 3, 5]).

The Matlab implementation of the algorithms used in this article as well as the database DB_{eval} can be found online: `http://www.benoit-fuentes.fr/publications`.

A. THEORETICAL INTRODUCTION OF BRAKES

In PLCA, observations $\boldsymbol{V}$ are the outcome of a generative process, and equations (2) and (3) result from the EM algorithm applied to this process. In order to formally introduce the brakes in equations (4) and (5), while remaining in the framework of the EM algorithm, one needs to slightly change the generative process of $\boldsymbol{V}$ (below, italics represent the addition of brakes to classic PLCA):

- $\forall (f,t) \in [\![1,F]\!] \times [\![1,T]\!]$, set $V_{ft} = 0$.
- Repeat J times:
  - draw (n,t) according to $P(n,t)$,
  - draw f according to $P(f|n)$,
  - set $V_{ft} = V_{ft} + 1$.
- *Repeat β^1_{brake} times:*

- *draw (n^0, t^0) according to $P(n^0, t^0)$ and do nothing with those variables.*

- *For each n, repeat β^2_{brake} times:*
  - *draw f^n according to $P(f^n|n)$ and do nothing with this variable.*

In this last step, n is no longer a random variable. The notation f^n (r.v. representing a frequency) is used because different r.v. names are needed depending on the value of n. β^1_{brake} and β^2_{brake} are integer values defining the strength of two brakes: one on parameters $P(n,t)$ and one on parameters $P(f|n)$. The main difference with classic PLCA is that some additional "virtual" variables linked to no observation are drawn, thereby making the estimation algorithm slower. We can now derive the EM algorithm. Firstly, the joint log-probability of hidden and observed variables is calculated (we denote $\bar{x}$ the set of all drawn variables x):

$$\begin{aligned}
&\mathcal{L}_\Lambda\left(\bar{f},\bar{t},\bar{n},\bar{n^0},\bar{t^0}\bar{f^1},\cdots,\bar{f^N}\right) = \\
&\ln\left(P\left(\bar{f},\bar{t},\bar{n},\bar{n^0},\bar{t^0}\bar{f^1},\cdots,\bar{f^N}\right)\right) = \\
&\sum_{j=1}^{J}\ln\left(P\left(n_j,t_j\right)\right)+\ln\left(P\left(f_j|n_j\right)\right) \\
&\quad+\sum_{j=1}^{\beta^1_{\text{brake}}}\ln\left(P\left(n_j^0,t_j^0\right)\right)+\sum_n\sum_{j=1}^{\beta^2_{\text{brake}}}\ln\left(P\left(f_j^n|n\right)\right). \quad (6)
\end{aligned}$$

Then, if we remember that V_{ft} corresponds to the number of times (f,t) is observed, the conditional expectation

$$Q_\Lambda = \mathbb{E}\left[\mathcal{L}_\Lambda\left(\bar{f},\bar{t},\bar{n},\bar{n^0},\bar{t^0}\bar{f^1},\cdots,\bar{f^N}\right)|\bar{f},\bar{t};\Lambda\right]$$

is given by:

$$\begin{aligned}
Q_\Lambda = &\sum_{f,t}\sum_n V_{ft}P\left(n|f,t\right)\left[\ln\left(P(n,t)\right)+\ln\left(P(f|n)\right)\right] \\
&+\quad\beta^1_{\text{brake}}\sum_{n,t}P(n,t|-)\ln\left(P(n,t)\right) \\
&+\quad\beta^2_{\text{brake}}\sum_n\sum_f P^n(f|-)\ln\left(P(f|n)\right). \quad (7)
\end{aligned}$$

The notation $P(x|-)$ is used to mean that the probability is an *a posteriori* probability but that the hidden variable x depends on no observation. Moreover, notation $P^n(f|-)$ means that the probability depends on the value of n.

In the E-step, posterior probabilities are calculated with respect to the current values of the parameters due to Bayes' theorem:

$$P(n|f,t) = \frac{P(n,t)P(f|n)}{P(f,t)}, \quad (8)$$

$$P(n,t|-) = P(n,t), \quad (9)$$

$$P^n(f|-) = P(f|n), \quad (10)$$

where $P(f,t)$ is given by equation (1). In the M-step, Q_Λ is maximized with respect to the parameters, under the constraint that all probability distributions sum to one:

$$P(n,t) \propto \sum_f V_{ft}P(n|f,t) + \beta^1_{\text{brake}}P(n,t|-), \quad (11)$$

$$P(f|n) \propto \sum_t V_{ft}P(n|f,t) + \beta^2_{\text{brake}}P^n(f|-). \quad (12)$$

By merging the E and M steps, one can deduce the multiplicative updates given by equations (4) and (5). It can be noticed that if we multiply $\boldsymbol{V}$, β^1_{brake} and β^2_{brake} by a same positive scalar, the update rules do not change. It is thus unnecessary to fix β^1_{brake} and β^2_{brake} to integer values, as well as it is not necessary to multiply an input TFR$^+$ $\boldsymbol{V}$ by a scaling factor so that its coefficients are integers.

REFERENCES

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

[2] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, F Nielsen and R Bhatia, Eds. Springer, 2012.

[3] T. Virtanen, A.T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. of ICASSP*, Las Vegas, NV, USA, 2008, pp. 1825–1828.

[4] M.V. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic Latent Variable Models as Nonnegative Factorizations," *Computational intelligence and neuroscience*, vol. 2008, no. 4, pp. 1–8, 2008.

[5] K.Y. Yilmaz, A.T. Cemgil, and U. Simsekli, "Generalized Coupled Tensor Factorization," in *NIPS*, Granada, Spain, 2011.

[6] B. Fuentes, R. Badeau, and G. Richard, "Harmonic Adaptive Latent Component Analysis of Audio and Application to Music Transcription," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 9, pp. 1854–1866, 2013.

[7] 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.

[8] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society.*, vol. 39, no. 1, pp. 1–38, 1977.

[9] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. of MLSP*, Kittilä, Finland, 2010, pp. 283–288.

[10] V. Emiya, R. Badeau, and B. David, "Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle," *IEEE Trans. on Audio, Speech, and Language Processing.*, vol. 18, no. 6, pp. 1643–1654, 2010.

[11] 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. of ICASSP*, Honolulu, Hawaii, USA, 2007, pp. 65–68.

[12] S.A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. of ISMIR*, Vienna, Austria, 2007, pp. 381–386.

[13] B. Fuentes, R. Badeau, and G. Richard, "Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation," in *Proc. of EUSIPCO*, Bucharest, Romania, 2012, pp. 2654–2658.