MULTIPITCH ESTIMATION USING A PLCA-BASED MODEL: IMPACT OF PARTIAL USER ANNOTATION

Camila de Andrade Scatolini, Gaël Richard, Benoit Fuentes

Institut Mines-Telecom, Télécom ParisTech, LTCI-CNRS
37-39, rue Dareau - 75014 Paris - France

ABSTRACT

In this paper we investigate the merit of partial user annotation for music transcription using a PLCA-based model. The original algorithm, called Blind Harmonic Adaptive Decomposition (BHAD), provides an estimation of the polyphonic pitch content of the input signal in an entirely unsupervised manner. In this paper, we have studied how the performances of the BHAD algorithm could be further improved by involving an user by means of a partial annotation. This user input allows for a better model initialisation with adapted or learned spectral envelope models. Furthermore, it is studied how a fine control of the convergence rate of some parameters can better exploit this additional information. It is then shown that this partial annotation can bring an improvement of up to 3% on the transcription of the remaining file.

***Index Terms*—** Multipitch estimation, PLCA, CQT, Semi-guided music transcription

1. INTRODUCTION

Multipitch estimation in musical recordings has received a great deal of attention in the last decade. This estimation task is known to be particularly challenging since a polyphonic music signal is typically composed of the superposition of the sound waves produced by all instruments in the recording. In the context of music transcription, multipitch extraction corresponds to the frame-by-frame estimation of all fundamental frequencies present in an audio signal. Two main issues arise in this estimation process : the superposition of the different partials of each note played simultaneously can lead to ambiguity in the case of harmonically related sounds; and the total number of notes being played simultaneously is unknown. Despite these difficulties, automatic multipitch estimation has many potential applications, including main melody extraction [1, 2, 3], cover song identification [4] (detecting whether two recordings are different renditions of the same musical piece), or more generally music transcription [5].

Several methods were proposed in the literature for multipitch estimation [6]. A number of approaches follows an iterative estimation strategy [7] while others will aim at jointly estimating all fundamental frequencies [8, 9, 10]. One class of techniques, which receives a sustained interest, exploits factorization models to represent the time-frequency representation of the audio signal as a sum of basic elements, called atoms.

A popular example of such decompositions is the non-negative matrix factorization (NMF), proposed by [11], and widely used in music analysis [12, 13, 2, 14]. In a probabilistic framework, Probabilistic Latent Component Analysis (PLCA) was shown to be particularly promising for audio signal analysis [15, 16, 17]. PLCA is a probabilistic tool for non-negative data decomposition, where the time-frequency representation of an audio signal (e.g. the spectrogram $P(f,t)$) is modeled as the histogram of J independent random variables $(f_j, t_j) \in [1, F] \times [1, T]$ distributed according to $P(f,t)$ where f and t respectively stands for frequency and time.

An algorithm called Blind Harmonic Adaptive Decomposition (BHAD) and its inherent model were presented in [10, 18] for better modelling real music signals. In this model, each musical note may present fundamental frequency and spectral envelope variations across repetitions. BHAD is an efficient algorithm and has obtained very good performances in an international evaluation campaign (ranked 2$^{\text{nd}}$ in the MIREX-12 Multiple Fundamental Frequency Estimation & Tracking task [19]). The original approach is entirely unsupervised as most approaches in this framework which rely on prior generic information or signal models to obtain a semantically meaningful decomposition. It was shown in some controlled cases that improved performances can be obtained by integrating a learning stage or by adapting the pre-learned models using a multi-stage transcription strategy (see [20] for example) or by involving a user during the transcription process [21], [22], [23].

In this paper, we have studied how the performances of the BHAD algorithm could be further improved by involving a user by means of a partial annotation (e.g. the user provides the transcription for the first ten seconds of the excerpt). This partial annotation allows for a better model initialisation with adapted or learned spectral envelope models. Furthermore, it is studied how a fine control of the convergence rate of some parameters can better exploit this additional information. It is then shown that this partial annotation can bring an improvement of up to 3% on the transcription of the remaining file.

The paper is organised as follows. In the following section, we recall the main concepts of the original BHAD model. Three strategies for semi-guided transcription are then presented in section 3. Experiments and results are given and discussed in section 4 and some conclusions are suggested in section 5.

2. ORIGINAL BHAD MODEL

The BHAD model is briefly described in this section. The reader is referred to [10] for further details. The absolute value of the normalized constant-Q transform (CQT) of a signal is modeled as a probability distribution $P(f,t)$. By introducing a latent variable c, the signal is first decomposed as the sum of a polyphonic harmonic signal ($c = h$) and a noise signal ($c = n$):

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

the notation $P_h(.)$ and $P_n(.)$ being used for $P(.|c = h)$ and $P(.|c = n)$. We recall below the main concepts of the harmonic signal model. However, since the model of the noise component is left unchanged in this work, the reader is referred to [10] for further details.

2.1. Polyphonic harmonic signal

At time t, the polyphonic component $P_h(f, t)$ is modeled as a weighted sum of different harmonic spectra, each one having its own spectral envelope and fundamental frequency $i \in [0, I-1]$. As the number of active notes is unknown, all possible fundamental frequencies are considered, with possibly zero weights:

$$P_h(f, t) = \sum_i P_h(i, t) P_h(f|i, t). \quad (2)$$

$P_h(i, t)$ and $P_h(f|i, t)$ respectively represent the energy and the normalized harmonic spectra of pitch i at time t. We further model $P_h(f|i, t)$ as a linear combination of Z fixed narrow-band harmonic kernels, sharing the same fundamental frequency i and having energy concentrated on the z^{th} harmonic:

$$P_h(f|i, t) = \sum_z P_h(z|i, t) P_h(f|z, i) \quad (3)$$

$$P_h(f|i, t) = \sum_z P_h(z|i, t) P_h(f - i|z). \quad (4)$$

In this last equation, an essentiel property of the CQT is exploited: a pitch modulation can be seen as a frequency shifting of the partials, and the kernel $P_h(f|z, i)$ can be deduced from a single template $P_h(\mu|z)$. All parameters, except for the fixed kernels $P_h(\mu|z)$, are estimated due to the Expectation-Maximization (EM) algorithm. By mean of a threshold applied on the time-frequency activations of harmonic spectra $P_h(i, t)$, it is then possible to estimate a MIDI pitch activations $\hat{A}(n, t)$ (n representing MIDI notes) and thus address the problem of multipitch estimation.

2.2. Initialisation and priors

Relevant initialisation can be seen as adding prior knowledge since parameters will likely converge toward a local optimum close to the initialisation. In [18] it is proposed that the spectral envelope coefficients $P_h(z|i, t)$ for each pitch i and time t are initialised as a descending slope in z, as often for musical instruments an energy decay of the partials in function of their frequency is observed.

In order to even better account for relevant initialization, it is also possible to use a "brake" on the well initialized parameters [24] (in our case the spectral envelopes $P_h(z|i, t)$). By slowing down their convergence rate during EM algorithm, it is more likely that their values after convergence are close to their initialisation.

In addition to relevant parameters initialisation, priors are used to integrate knowledge about the nature of the signals. A resemblance prior [18] is applied to $P_h(z|i, t)$ for each frequency component i which allows to take into account that the spectral envelope slowly evolves over time. A sparseness prior is applied to the time-frequency activations $P_h(i, t)$ which helps to model the signal with the least amount of notes.

3. STRATEGIES FOR SEMI-GUIDED TRANSCRIPTION

It is expected that involving the user in the transcription process should improve the overall transcription performances. In this paper, we evaluate a rather simple interaction process where the user has manually transcribed beforehand the first ten seconds of each processed musical recording. This partial transcription is then used to create initial templates of the spectral envelope $P(z|i, t)$ for each note i. If a note appears more than once in these first ten seconds, the template is obtained by averaging all occurrences of this given note for each kernel $z \in [1, Z]$. The transcription algorithm BHAD is then run as in the unsupervised case but with an improved initial estimation of, at least, some of the spectral envelopes $P(z|i, t)$.

3.1. Strategies for initialisation of the non-annotated notes

Unless if the music is highly repetitive, the note present in the first ten seconds only represent a subset of all notes played in the musical recording. The envelope templates of the remaining notes cannot be learned but still have to be initialised to some value before the BHAD algorithm is run. We evaluate below three strategies for initialising the templates of the remaining notes:

1. Keep the slope initialisation as in the original BHAD model: in this case we just consider the templates if they can be obtained from the learning phase, otherwise the spectral coefficients are initialised as a descending slope in z;

2. Copy the previous note template: the template of a given note i_n is repeated for all subsequent notes $i \in [i_{n+1}, i_{p-1}]$ until another template is found (note i_p);

3. Interpolate neighbour notes' templates: the template for the notes $i \in [i_{n+1}, i_{p-1}]$ are obtained by linearly interpolating the templates of the notes i_n and i_p (on a dB scale).

3.2. Strategies for controlling the convergence rate of the annotated notes vs. non-annotated notes

The goal of the learning phase using the partial user annotation is to improve the initial estimation of the spectral envelopes. If envelope parameters are well initialised, it is probable that their value are not very far from the ideal envelope templates. It is then desirable to exploit this information to control the convergence rate of this parameter compared to other parameters of the models which are less well initialised. With BHAD, the convergence rate is controlled using the concept of brake (see section 2.2 or [10]). In the original BHAD model the convergence rate coefficient is equally applied to all notes. We suggest a modification of the BHAD model where the convergence rate coefficient $\beta_{brake}(n)$ depends on the notes, so we can apply a stronger brake for the templates of the notes which were actually learned:

$$\beta_{brake}(n) = \begin{cases} \beta_1, & \text{if } i_n \in \text{learning base} \\ \beta_0, & else \end{cases}, \text{ with } \beta_1 > \beta_0. \quad (5)$$

4. EXPERIMENTS AND RESULTS

4.1. Database and evaluation metrics

To assess the quality of the multipitch estimation provided by the algorithm, its estimated activations $\hat{A}(n,t)$ are compared with the ground truth $A(n,t)$ and then evaluated in terms of F-measure. This measure combines the precision and recall measures in order to give a global index of the estimation's quality:

$$\mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}}.$$

The precision $\mathcal{P}$ indicates the ratio of activations correctly estimated by the total activations estimated while the recall $\mathcal{R}$ indicates the ratio of correctly estimated activations by all activations in the ground truth:

$$\mathcal{P} = \frac{\sum_{n,t}\hat{A}(n,t)A(n,t)}{\sum_{n,t}\hat{A}(n,t)}, \quad \mathcal{R} = \frac{\sum_{n,t}\hat{A}(n,t)A(n,t)}{\sum_{n,t}A(n,t)}.$$

We used the QUASI-Transcription database elaborated under the QUAERO[1] project and [18]. It contains pieces of contemporary music, belonging to different genres with a high degree of polyphony (see table 1 which details the characteristics of each song in the database). The instruments are a mixture of synthesis and acoustic instruments. For virtual instruments, transcripts were obtained from the corresponding MIDI file, while for acoustic instruments, they were manually transcribed. The database is split in two : a training database built from the first 10s of each file and a test database which gathers the remaining part of all songs. All results are given on the test database.

Song name	Duration	# inst./notes	Poly. (mean/max)
RockSong	01'14"	9/1039	3.9/10
Choir	01'11"	4/224	3.4/4
Filter	01'19"	19/2418	5.9/11
Unison	01'17"	6/561	5.6/9
Accelerando	01'49"	3/1046	2.3/6

Table 1. QUASI-Transcription database

4.2. Experiments and Results

We compared the algorithm performances for the unsupervised and semi-guided approaches for all combination of priors (sparseness - S and resemblance - R) and guided parameters convergence (brake - B), including the case where no prior or brake was used (NO). The initialisation for the unsupervised approach is the one that is proposed in [18]: for each pitch i and time t the spectral envelope is initialised as a descending slope in z. In the semi-guided approach, the spectral envelopes are initialised either by the templates obtained in the learning phase or one of the three strategies of initialisation for the non-annotated notes (slope, copy and interpolation).

The results of these tests are shown in Figure 1, where the continuous lines correspond to the unsupervised approach and the lines with symbols correspond to the user-guided approach, and Table 2. The table gathers the results obtained a posteriori, that is with a choice of the detection threshold that maximises the algorithm's performance for each combination of priors and brake on the entire test database. The best results for the options without the brake (e.g. options NO, S, R and SR), obtained with the semi-guided approach with slope initialisation of the non-annotated notes are presented in Figure 1 for a range of detection threshold values.

[1] http://www.quaero.org

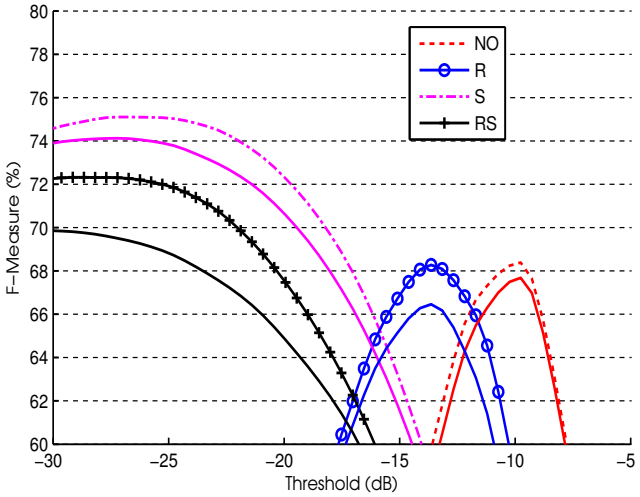

Fig. 1. Comparison of the mean F-Measure for all 5 songs in the database in function of the threshold for the semi-guided approach with slope initialisation for the non-annotated notes (symbols) and the unsupervised approach (continuous line) for all combinations of priors without brake coefficients.

Options	**Unsupervised**	**Semi-Guided**		
		Slope	Copy	Interpolation
NO	67.68	**68.38**	65.00	43.68
S	73.91	**75.07**	71.44	52.07
R	66.46	**68.28**	64.13	47.05
RS	69.51	**72.30**	67.60	52.35
B	77.85	**78.63**	75.88	50.90
SB	74.64	**76.54**	72.47	45.68
RB	77.39	**79.49**	73.89	50.79
RSB	74.85	**77.02**	70.52	47.00

Table 2. Mean F-Measure for the semi-guided and unsupervised approaches, considering the different initialisations of the non-annotated notes (Slope, Copy and Interpolation) and the different options of the BHAD algorithm : sparseness (S) and resemblance (R) priors, convergence rate parameter (brake B); Option NO refers to no prior and no brake.

The results in Table 2 show that the use of the semi-guided approach with slope initialisation of the non-annotated notes increases the mean F-measure between 1% and 3%. This table also shows that, for the copy initialisation the F-measure decreases around 2% or even 4% when compared to the unsupervised approach. This degradation of the results is even more important in the case of the interpolated initialisation. We note that, in general, the gain in terms of F-measure is greater for the case where the brakes are used for

guiding the convergence of the parameters.

Using copied spectral envelope templates as initialisation for non-annotated notes introduces a bias in the estimation task by assuming that all the notes are present. As the algorithm starts with a calculated spectral envelope instead of a simple slope in z, it causes the algorithm to assume that these notes are present (smaller precision when compared to the unsupervised approach). Templates obtained by interpolating adjacent notes presented the worst results since in this case besides introducing the copy initialisation bias, the templates were not really learned as in the previous case.

Regarding the different strategies for controlling the convergence rate of the annotated notes, we tested two cases:

1. The brake coefficient is only applied to annotated notes ($\beta_0 = 0$);
2. The same brake coefficient used in the previous tests ($\beta_0 = 10$) is applied to the non-annotated notes but a greater coefficient is applied to the annotated notes.

Similarly to the previous experiment, the results are given in Table 3 for all options with a choice *a posteriori* of the detection threshold. The best results considering a different convergence rate of the calculated spectral envelope templates ($\beta_0 = 10$ and $\beta_1 = 20$) and slope initialisation of the non-annotated notes for the semi-guided approach are presented in Figure 2.

Option	**No brake**	**Brake annotated notes** ($\beta_0 = 0$)		
	$\beta_{0,1} = 0$	$\beta_1 = 0.1$	$\beta_1 = 1$	$\beta_1 = 10$
B	**68,38**	68,25	66,43	62,93
SB	**75,07**	74,84	74,29	72,74
RB	68,28	68,53	**69,20**	69,16
RSB	72,30	72,35	72,53	**72,66**

(a) Brake coefficient only for annotated notes.

Option	**Even brake**	**Brake all notes** ($\beta_0 = 10$)		
	$\beta_{0,1} = 10$	$\beta_1 = 10.1$	$\beta_1 = 11$	$\beta_1 = 20$
B	78,63	78,63	**78,66**	**78,66**
BS	76,54	76,54	**76,55**	**76,55**
BR	79,49	79,49	79,51	**79,61**
BSR	77,02	77,02	**77,03**	76,98

(b) Brake coefficient greater for annotated notes.

Table 3. Mean F-Measure considering the different strategies for controlling the convergence rate of the annotated notes and the different options of the BHAD algorithm (priors).

The results in Table 3.(a) show that when we apply the brake coefficient only for the annotated notes, the improvement of the F-measure depends on the considered priors. In particular, when the resemblance prior is used we note that the results improve up to 1% by using the brake coefficient for the annotated notes. However, the results in Table 3.(b) (where the brake coefficient is used for all notes) show that there is little interest to use a more restrictive convergence rate for annotated notes compared to the one used for non-annotated notes (a maximum of 0.1% increase in F-measure is observed).

5. CONCLUSION

As highlighted in an international evaluation campaign, the original BHAD algorithm is a powerful approach for multipitch extraction of music recordings. Since it is entirely unsupervised, it relies on a very generic model for the initial estimation of the notes spectral templates. It is shown in this paper that a substantial gain in performance can be obtained with the introduction of a learning step with a partial user annotation of even a small part of the song (first 10 seconds). This partial annotation allows for a better initialisation of the model parameters by using learned spectral envelope models for the notes that are present in the annotation. The experiments have shown that this partial annotation can lead to an improvement of up to 3% in terms of F-measure in a task of multipitch estimation. It was also shown that a fine control of the convergence rate of these learned models helps to better exploit this additional information.

Future work will be dedicated to the investigation of alternative strategies for optimizing the user annotation effort, for example by rather annotating specific parts of the song such as the main melody, the bass line or one of the chorus of the song. Another strategy would involve the user in a more iterative way for example letting the user transcribe the segments where the algorithm is the least confident.

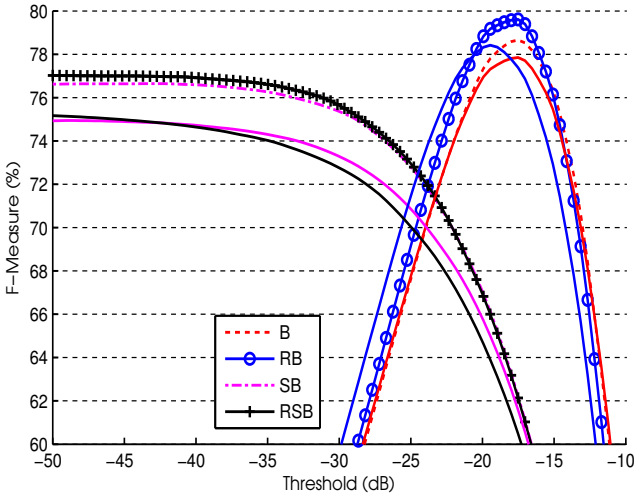

Fig. 2. Mean F-Measures as a function of the threshold for the semi-guided approach with slope initialisation for the non-annotated notes (symbols) and the unsupervised approach (continuous line) for all combinations of priors with brake coefficients with $\beta_1 = 20$ for annotated notes and $\beta_0 = 10$ elsewhere.

6. REFERENCES

[1] J. Salomon, E. Gomez, D. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications and challenges," *IEEE Signal Processing magazine*, vol. 31, no. 2, pp. 118–134, Mar. 2014.

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

[3] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 425–428.

[4] J. Salamon, J. Serrà, and E. Gómez, "Tonal representations for music retrieval: From version identification to query-by-humming," *Internacional Journal of Multimedia Information Retrieval, special issue on Hybrid Music Informantion Retrieval*, vol. 2, no. 1, pp. 45–58, Mar. 2013.

[5] S.W. Hainsworth, *Techniques for the Automated Analysis of Musical Audio*, Ph.D. thesis, University of Cambridge, UK, 2004.

[6] M.G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2009.

[7] A Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255–266, Feb 2008.

[8] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, 2006.

[9] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, March 2010.

[10] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, August 2012, pp. 2654–2658.

[11] 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, Oct. 1999.

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

[13] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 1825–1828.

[14] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *11th International Society for Music Information Retrieval Conference*, 2010, pp. 489–494.

[15] E. Benetos and S. Dixon, "Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1111–1123, Oct. 2011.

[16] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Workshop on Advances in Models for Acoustic Processing at NIPS*, 2006.

[17] P. Smaragdis, B. Raj, and M. V. S. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 2069–2072.

[18] B. Fuentes, *L'analyse probabiliste en composantes latentes et ses adaptations aux signaux musicaux. Application à la transcription automatique de musique et à la séparation de sources.*, Ph.D. thesis, Télécom ParisTech, Paris, France, 2013.

[19] MIREX'2012 Evaluation, "Multiple fundamental frequency estimation - task 2: Note tracking," http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results.

[20] E. Benetos, R. Badeau, T. Weyde, and G. Richard, "Template adaptation for improving automatic music transcriptio," in *International Society for Music Information Retrieval (ISMIR)*, October 2014.

[21] H. Kirchhoff, S. Dixon, and A Klapuri, "Missing template estimation for user-assisted music transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, May 2013, pp. 26–30.

[22] 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, pp. 1118–1133, May 2012.

[23] N. J. Bryan, G. J. Mysore, and G. Wang, "Source separation of polyphonic music with interactive user-feedback on a piano roll display," in *International Society for Music Information Retrieval (ISMIR)*, November 4-8 2013.

[24] B. Fuentes, R. Badeau, and G. Richard, "Controlling the convergence rate to help parameter estimation in a plca-based model," in *European Signal Processing Conference (EUSIPCO)*, Lisbon, September 2014.