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

Camila de Andrade Scatolini, Gaël Richard

Télécom ParisTech
37-39, rue Dareau - 75014 Paris - France
andrade@telecom-paristech.fr, gael.richard@telecom-paristech.fr

ABSTRACT

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

1. INTRODUCTION

Music's sound environment is a complex result of the combination of multiple sound sources. For some tasks such as melody extraction, each musical instrument in a recording can be considered as one source. However, for tasks such as music transcription one can see each musical note as a different source. The sound of most instruments being a quasi-periodic signal and therefore formed by frequencies that are multiple of a fundamental frequency F_0, it is relevant to estimate this frequency, which is one of the major attributes that characterize a musical note.

Multipitch estimation is the frame-by-frame estimation of all fundamental frequencies present in an audio signal. As multiple notes can be played simultaneously, the signal contains several F_0 at a given moment. Two main issues arise in this kind of estimation: the superposition of the different partials can lead to ambiguity in the case of sounds with an octave relationship and the total number of sources, i.e. notes, being played simultaneously is unknown.

Several methods for multipitch estimation were proposed in the literature. One class of techniques, in particular, models a transformation of the audio signal as a sum of basic elements, called atoms, which are related to the physic signal and carry symbolic information. A popular example of this kind of decomposition is the non-negative matrix factorization (NMF), proposed by [1], and widely used in music analysis [2, 3].

Other methods [4, 5] address this decomposition as a statistic inference problem rather than analytic optimization. Probabilistic Latent Component Analysis (PLCA) [6, 7] is a probabilistic tool for non-negative data decomposition, where the time-frequency representation of an audio signal is modeled as the histogram of J independent random variables $(f_j, t_j) \in [1, F] \times [1, T]$ distributed according to $P(f, t)$.

An algorithm called Blind Harmonic Adaptive Decomposition (BHAD) and its inherent model were presented in [8, 9] for better modelling real music signals, where the notes present fundamental frequency and spectral envelope variations. It uses the absolute value of the CQT of an audio signal as the time-frequency representation and its decompositions and parameters are found using the Expectation-Maximization (EM) algorithm.

Even though it has rather good performances (2$^{\text{nd}}$ place in the MIREX-12 Multiple Fundamental Frequency Estimation & Tracking task), we have studied how its performances could be improved by using a better model initialisation with learned spectral envelope models. Also, it's desirable to assess the impact of controlling the convergence rate of this parameter in order to better take into account this initialisation.

Main contributions compared to BHAD:

1. Impact of better initialisation of spectral templates using a partial annotation of the music file

2. Interest of controlled convergence rate of the spectral templates (e.g. brake parameters)

This paper briefly describes the original BHAD model (section 2) and presents some strategies for semi-guided transcription (section 3). In the section 4 we present our experiments and results and finally the section 5 concludes this paper.

2. ORIGINAL BHAD MODEL

A latent variable c is introduced in order to decompose the absolute value of the CQT V_{ft} of a music signal as the sum of a polyphonic harmonic signal ($c = h$) and a noise signal ($c = n$). The notation $P_h(.)$ and $P_n(.)$ is used for $P(.|c = h)$ and $P(.|c = n)$, respectively. The time-frequency distribution of the music signal is:

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

2.1. Noise signal

The noise signal is modeled as the convoltion of a narrow-band window and a noise time-frequency distribution:

$$P_b(f,t) = \sum_i P_b(i,t)P_b(f-i). \tag{2}$$

2.2. Polyphonic harmonic signal

For a given instant t, the polyphonic component of the signal $P_h(f,t)$ is a weighted sum of the different harmonic spectra, each with its own spectral envelope and its own fundamental frequency $i \in [0, I-1]$. As we don't want to limit the number of notes present, we consider all the possible fundamental frequencies with probably zero weights:

$$P_h(f,t) = \sum_i P_h(i,t)P_h(f|i,t), \tag{3}$$

$P_h(i,t)$ is the note's energy at time t and $P_h(f|i,t)$ is its normalized harmonic spectrum. Each harmonic spectrum is decomposed as a linear combination of Z narrow-band harmonic kernels sharing the same fundamental frequency i and having energy concentrated on the z^{th} harmonic:

$$\begin{array}{ll} P_h(f|i,t) & = \sum_z P_h(z|i,t)P_h(f|z,i) \\ & = \sum_z P_h(z|i,t)P_h(f-i|z) \end{array} \tag{4}$$

In this last equation it's used the main advantage of working with the CQT: a pitch modulation can be seen as a frequency shifting of the partials: $P_h(f|z,i) = P_h(f-i|z)$.

Finally, the complete equation for the BHAD model is written as:

$$\begin{array}{ll} P(f,t) = & P(c=h)\sum_{i,z} P_h(i,t)P_h(z|i,t)P_h(f-i|z) \\ & +P(c=n)\sum_i P_n(i,t)P_n(f-i) \end{array} \tag{5}$$

2.3. Initialisation and priors

We want to estimate the following parameters: $\Lambda = \{P(c), P_h(i,t), P_h(z|i,t), P_n(i,t)\}$. Since this is an iterative optimization algorithm, their initialisation plays a very important role in the quality of the decomposition. The spectral envelope coefficients $P_h(z|i,t)$ are initialised as a descending slope in z, since often for musical instruments it is observe an energy decay of the partials in function of their frequency.

In addition to well-initialise the settings, a way to help the algorithm to converge to relevant solutions is the use of priors, which allow integration of knowledge about the nature of the signals. A resemblance prior is applied to $P_h(z|i,t)$ for each frequency component i. It supposes that the spectral envelope varies little over time. A sparseness prior is applied to the time-frequency activations $P_h(i,t)$ and it supposes that the best solution is the one that models the signal with the least amount of notes.

Another concept is introduced in [9] in order to guide the convergence of the parameters. A "brake" is applied to $P_h(z|i,t)$ and it considers that its true value is not far from the initialisation.

3. STRATEGIES FOR SEMI-GUIDED TRANSCRIPTION

Different strategies for the semi-guided transcription were taken into account during the experiments. We considered the manual annotation of the first 10 seconds of a music file. Learning is achieved by running the algorithm with fixed time-frequency activations (annotated notes). The template for spectral envelope coefficients for each harmonic component i is obtained by averaging over time all occurrences of each note.

Strategies for initialisation of the non-annotated notes

Not all notes of the considered note range are present in the learning phase, therefore it's necessary to define how the non-annotated notes will be initialised. Three possibilities were tested:

1. Keep the slope initialisation as in the original BHAD model;
2. Copy the previous note template;
3. Interpolate neighbour notes' templates.

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

As the goal is to improve the initialisation of spectral envelope coefficients, we suppose that their values are rather close to the final values thus it's desirable to apply a stronger "brake" in these learned envelopes. In the original BHAD model the convergence rate coefficient is equally applied to all notes. However, even tough we have realized a better initialisation for the notes present in the learning phase, the other notes may not be as well initialised, for example, if we keep slope initialisation. In order to cope with this problem we suggest a modification of the BHAD model where the "brake" coefficient $\beta_{brake}(n)$ depends on the notes:

$$\beta_{brake}(n) = \left\{ \begin{array}{ll} \beta_1, & \text{if } n \in \text{learning base} \\ \beta_0, & else \end{array} \right. , \text{ with } \beta_1 > \beta_0. \tag{6}$$

4. EXPERIMENTS AND RESULTS

4.1. Database and evaluation metrics

To assess the quality of the multipitch estimation provided by the algorithm it is necessary to have an evaluation database containing musical recordings and their corresponding ground truth. The algorithm's estimates are compared with the ground truth and then we use measures to evaluate the quality of the resulting estimation.

$\hat{A}(n,t)$ are the estimated activations for a given signal and $A(n,t)$ the corresponding ground truth. The quality of this estimation is evaluated using 3 conventional measures: the precision indicates the ratio of pitchs correctly estimated by the total pitchs estimated:

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

The recall indicates the ratio of correctly estimated pitchs by all pitchs in the ground truth:

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

Finally, the F-Meesure combines both measures in order to give a global measure of the quality of the estimation:

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

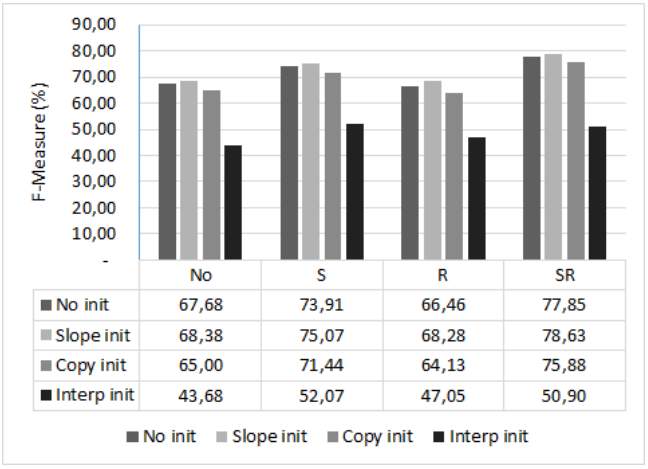
We used the QUASI-Transcription database elaborated under the QUAERO[1] project and [9]. It contains pieces of contemporary music, belonging to different genres. The instruments are a mixture of synthesis and acoustic instruments. For virtual instruments, transcripts were obtained from the corresponding MIDI file, while for acoustic instruments, they were made by hand. The table 1 presents the characteristics of each song in the 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 songs description

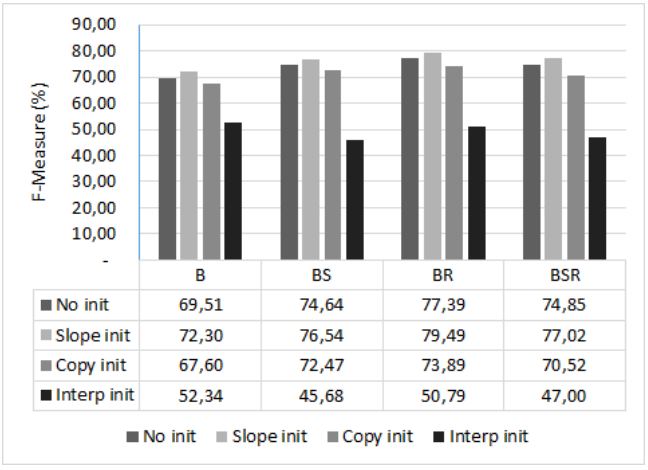
4.2. Experiments and Results

We compared the algorithm performances with and without initialisation 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 user (NO). All three strategies for initialisation of the non-annotated notes were tested (slope, copy and interpolation). The results of these tests are shown in the figure 1. We picked the detection threshold which maximised the algorithm's performance for each combination of priors and parameters convergence rate.



	No	S	R	SR
No init	67,68	73,91	66,46	77,85
Slope init	68,38	75,07	68,28	78,63
Copy init	65,00	71,44	64,13	75,88
Interp init	43,68	52,07	47,05	50,90

(a) No brake options.



	B	BS	BR	BSR
No init	69,51	74,64	77,39	74,85
Slope init	72,30	76,54	79,49	77,02
Copy init	67,60	72,47	73,89	70,52
Interp init	52,34	45,68	50,79	47,00

(b) Brake options.

Fig. 1. Comparison of the mean F-Measure for all 5 songs in the database with and without initialisation, considering the different initialisations of the non-annotated notes (slope, copy and interpolation) and the different options of the BHAD algorithm (priors and parameters convergence rate).

The results in the figure 1 show that, for the slope initialisation the F-measure increases between 1% and 2% for the options without brake and between 2% and 3% for the options with brake. This figure also shows that, for the copy initialisation the F-measure decreases around 2% or even 4% when compared to the reference test (no initialisation). This degra-

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

dation 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, even in the case where the brake is applied uniformly to the notes annotated and non-annotated.

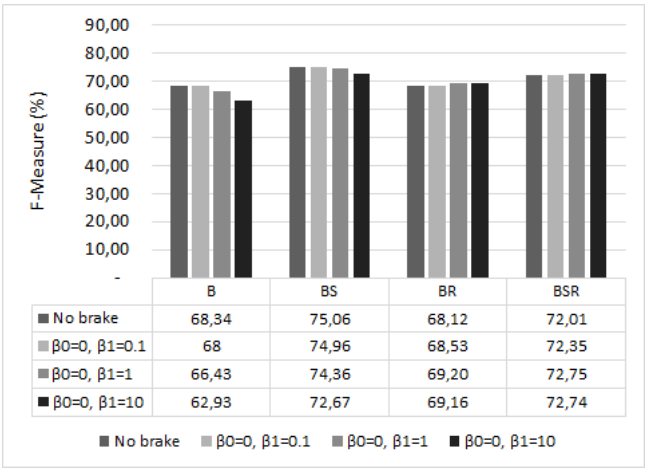
Copying spectral envelope coefficients for non-annotated notes introduces a bias in the estimation task by assuming that all the notes are present. As the algorithm starts with a good solution for spectral envelope coefficients, it causes the algorithm to say that these notes are present (smaller precision when compared to the reference test with no initialisation). Initialisation by interpolating adjacent notes presented the worst results since in this case besides introducing the copy initialisation bias, the templates weren't really learned as in the the copy initialisation 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; and $2)$ we use the same brake coefficient used in the previous tests and we apply a greater coefficient to the annotated notes. The results in the figure 2 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 we use the resemblance prior we note that the results improve up to 1% by using the brake coefficient in the annotated notes. However, if we already use a brake coefficient for annotated and non-annotated notes, the use of a more restrictive convergence rate for annotated notes has a smaller effect in the F-measure (around 0.5%).

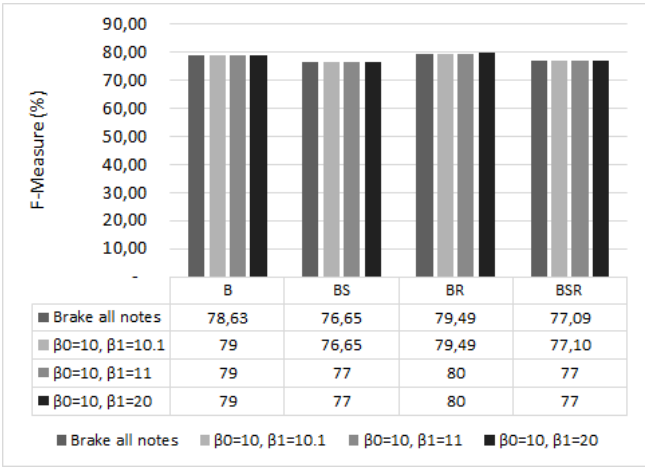
5. CONCLUSION

From the results in the section 4 we observe that even though BHAD algorithm is already efficient it is possible to further improve its performance by introducing a learning step with a partial user annotation.

From the results shown in the previous section we see that the algorithm can benefit from a preliminary training step with the annotation of even a small part of the song (the first 10 seconds).



(a) Brake coefficient only for annotated notes.



(b) Brake coefficient greater for annotated notes.

Fig. 2. Comparison of the mean F-Measure for all 5 songs in the database considering the different strategies for controlling the convergence rate of the annotated notes and the different options of the BHAD algorithm (priors).

6. REFERENCES

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

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

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

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

[5] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[6] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "A probabilistic latent variable model for acoustic modeling," in *In Workshop on Advances in Models for Acoustic Processing at NIPS*, 2006.

[7] Paris Smaragdis, Bhiksha Raj, and Madhusudana V. S. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, 2008, pp. 2069–2072.

[8] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. of EUSIPCO*, Bucharest, Romania, August 2012, pp. 2654–2658.

[9] Benoît 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.