Joint f0 and inharmoncity estimation using second order optimization

Henrik Hahn, Axel Röbel

Joint f0 and inharmoncity estimation using second order optimization

Henrik Hahn
IRCAM-CNRS-UPMC UMR 9912-STMS
henrik.hahn@ircam.fr

Axel Röbel
IRCAM-CNRS-UPMC UMR 9912-STMS
axel.roebel@ircam.fr

ABSTRACT

A new method is presented for the joint estimation of the inharmonicity coefficient and the fundamental frequency of inharmonic instrument sounds. The proposed method iteratively uses a peak selection algorithm and a joint parameters estimation method based on nonlinear optimization. We further introduce an adapted tessitura model to evaluate our proposed method for piano sounds and to compare it with state-of-the-art techniques.

1. INTRODUCTION

The stiffness of instrumental strings effectuates the frequencies of the modes of vibration to be highly inharmonic. This effect is decisive for most string based instruments and marks a significant part of the perceptive sound characteristic of the piano [1]. Inharmonicity means that the partial frequencies are not exact integer multiples of their fundamental but located at increased positions. The amount of increase is reflected by the inharmonicity coefficient β, while the frequency f of a partial k can be expressed for all partials K present in a signal by the relation:

$$f_k = k f_0 \sqrt{1 + k^2 \beta}, \quad k = 1 \dots K \tag{1}$$

where f_0 denotes the signals fundamental frequency, which is in fact a theoretical value, as there is no partial with that specific frequency present in an inharmonic signal. Hence, the inharmonicity coefficient β as well as the fundamental frequency f_0 can not easily be measured from an instruments signal, but they need to be taken into account for a lot of different applications, like f_0-estimation and harmonic sinusoidal analysis, as well as for prior knowledge to control sound synthesis of string based instruments. And finally, demixing of sound mixtures is an emerging topic, which also relies on good estimations of the inharmonicity coefficient and the fundamental frequency.

In the following section we give a brief overview on three previous estimation methods and point out several drawbacks of them in section 3 before we give a detailed description of our approach, which aims to solve these drawbacks. An extensive evaluation of our approach with an adapted tessitura model, comparing it with the three other methods is presented in the 4th section.

2. PREVIOUS METHODS

Several methods for the automatic estimation of the inharmonicity factor β with according refinement of the fundamental frequency f_0 have been proposed in the past years. Galembo and Askenfelt proposed a method [2] based on inharmonic comb filtering (ICF). In this method, the parameters for the inharmonic comb filter have been found by an exploration of a vast range of possible parameter values within three consecutive steps and refining the parameter grid in each. The algorithm finally interpolates the best parameter sets to obtain its f_0 and β-coefficient. Hodgkinson et al. proposed a method [3] using median-adjustive trajectories (MAT). This algorithm works in an iterative manner in which a partial k of the inharmonic series is selected and used for improving the estimate of β and f_0. The improved estimates are then used to search the next partial k. The most recent approach is based on Non-negative matrix factorization by Rigaud et al. [4, 5] aiming at the joint estimation of f_0 and β-coefficients for several fundamental frequencies at once with a specific focus on the polyphonic case. Another approach has been proposed in [6] showing similar accuracy, but improved computational performance to the ICF method.

3. PROPOSED METHOD

3.1 Drawbacks in recent methods

All recent methods we studied so far share similar drawbacks. First of all, they usually work with a fixed maximum of around 30 partials or fixed amplitude thresholds to avoid using too noisy signal components for the estimation. But, especially low pitched piano tones may exhibit very rich spectra containing more than 200 partials. For an analysis which tries to reliably identify as much partials as possible in such a signal, the estimation of the β coefficient needs to be executed for far more partials, because slight deviations in the β estimation will remain unnoticed. Figure 1 illustrates how such small errors in the estimation of the β coefficient result in misleading partial detection. Increasing the amount of partials for the estimation of β is by no means a trivial task as it requires a suitable strategy for selecting reasonable spectral peaks and rejecting noisy signal components. Furthermore, some approaches need at least 5 partials for a reliable estimation, but high pitched piano notes or moderately high pitched but with very low intensity do not contain more than 3 to 4 partials. Especially low intensity signals require a robust distinction between noise and sinusoid within a peak selection process but also require the estimation to be robust against noisy

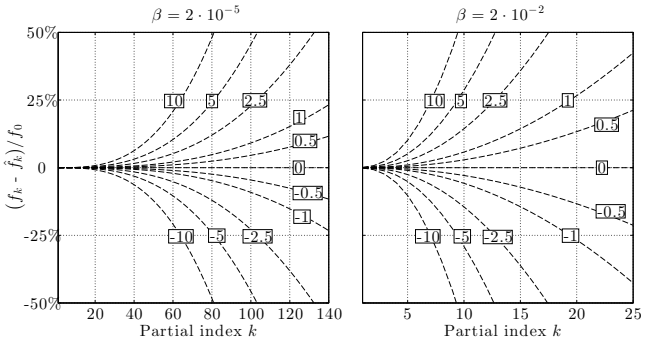

Figure 1. Analysis of the effect of the deviation in β. Boxed values indicate hypothetical deviations of the β value from its 'real' value in percent. Dashed curves demonstrate the resulting deviation in frequency estimation for respective partial index.

partials. Previous approaches often use heuristics to either neglect noisy partials during the peak selection or reduce their influence in the estimation process.

3.2 General method description

The proposed method estimates jointly the inharmonicity coefficient β and the fundamental frequency f_0 in an iterative manner, which can be used on several frames at once and is illustrated in figure 2. For the algorithm a signal segment $y(t)$ behind the signals attack point is selected to ensure, the algorithm analyses no transient components. A standard f_0 estimation [7] is applied and the f_0 information is then being used to set the analysis parameters for the STFT adaptively to guarantee suitable analysis window lengths according to the fundamental. The STFT is taken for N overlapping frames n yielding $Y(f, n)$ and all spectral bins are classified into the 3 classes: main lobe, side lobe or noise component using the *peak classification* method proposed by Zivanovic et al. [8]. The algorithms

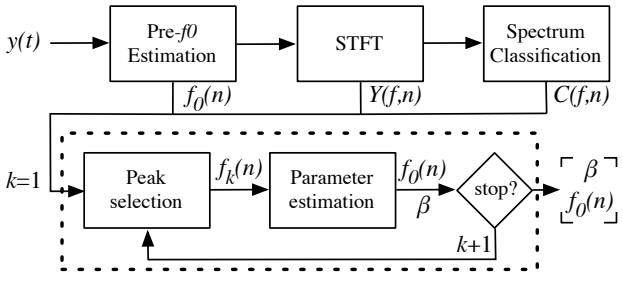

Figure 2. General scheme of proposed iterative method.

main loop identifies a valid peak for the current partial index within each frame and estimates a new f_0 for each frame n and a new β for all frames within each iteration until some abort criterion has been reached. With increasing partial index, the estimated parameters converge to our target values.

3.3 Peak selection step

The selection of a valid peak within the spectrum is done in 4 steps:

1. Estimate the frequency of the current partial $\hat{f}_k(n)$ by using eq. (1). Use the initial $f_0(n)$ and $\beta = 0$ for the first iteration, and the updated values in all later ones.

2. Select all spectral peaks classified as main lobe within a narrow band f_b around the estimated partials frequency $\hat{f}_k(n)$:

$$\hat{f}_k(n) - pf_0(n) \leq f_b \leq \hat{f}_k(n) + pf_0(n), p = .25$$

3. If two or more peak candidates have been found within at least one frame, we apply a logarithmic amplitude weighting function using a Hann window, centered at the estimated position $\hat{f}_k(n)$, with window length f_b and select the peak with the strongest logarithmic amplitude after weighting.

4. Refine the frequency of the selected peaks by QIFFT and bias correction as proposed by Abe et al. [9].

3.4 Estimation step

With at least 3 partials within one frame, we can estimate the parameters β and $f_0(n)$ for all frames n. As shown in eq. (2), we use the squared deviation of our estimated values from the measured partial frequencies normalized with the fundamental frequency to achieve equal error surface scalings for all possible fundamental frequencies. The final objective function with normalizations according to the number of frames N and amount of partials per frame $K(n)$ is given in eq. (3).

$$R = \frac{1}{2}\left(\frac{f_k(n) - kf_0(n)\sqrt{1+k^2\beta}}{f_0(n)}\right)^2 \quad (2)$$

$$O_1 = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K(n)}\sum_{k=1}^{K(n)} R \quad (3)$$

As the objective function (3) reflects the least–mean–squared (LMS) error of all f_0-normalized deviations of our partial frequency estimations with their measured peak frequency counterparts, optimization reflects a fitting of eq. (1) to the measured data in the LMS sense. The optimization is being done by a gradient descent approach, whereas we utilize the method of the scaled conjugate gradient [10], denoted CG throughout this document, for faster convergence compared with other methods. The gradient functions for both parameters are shown in eq. (4) and (5).

$$\frac{\partial R}{\partial \beta} = -\frac{k^3}{2\sqrt{1+k^2\beta}} \quad (4)$$

$$\frac{\partial R}{\partial f_0(n)} = -\frac{f_k(n)}{f_0(n)^2} \quad (5)$$

3.5 Stop criterion

We only use two disjunctive abort criteria: If the next partial $\hat{f}_k(n)$ in the peak selection process would raise above the Nyquist frequency within one frame n or if no valid partial has been found for 3 consecutive iterations in at least one frame of the main loop. This means, the algorithm tries to use as much partials as possible of the signal, since it only stops, if the signals maximum bandwidth or some supposed noise level has been reached.

4. EVALUATION

For the evaluation we will compare the results of our proposed method with the results of the 3 algorithms mentioned in chapter 2: ICF, MAT and NMF. Our proposed method will be denoted CG in the figures. We will use an artificial data sound of inharmonic sounds, created using an additive synthesis model and inharmonicity values taken from the tessitura model for the β coefficient shown in [5] as well as the 3 piano data sets from the RWC library [11] and a piano sound set taken from the IRCAM Solo Instruments library recorded with two microphones. The artificial data set will be used to compare all β coefficient estimation algorithms with a given ground truth. For the general evaluation of all data sets we will establish a tessitura model for the evolution of the coefficient for all sound samples contained in each data set. The tessitura model for the evolution of β over the MIDI index is derived from [5] and will be used to measure the variance of each estimation algorithm to quantify its accuracy. Furthermore, we will compare the computational efficiency of all algorithms by measuring their realtime factors. For each algorithm a MATLAB^{TM} implementation has been used therefore the realtime factors are more suitable for a comparison in between the algorithms rather than to give an indication for the performance of native implementations. For all algorithms we used equal analysis parameters to ensure all algorithms analyze exactly the same frames of the signals and as most other algorithms also need a pre-f_0 estimation, we used the same pre-f_0 for all of them. The window length for the STFT was set to 6 times the roughly estimated fundamental with 4 times spectral oversampling and a *blackman* window. As our algorithm works on several frames, we took 3 consecutive frames with a hopsize of 1/8 of the analysis window length, whereas the other algorithms analyzed the 3 frames independently.

4.1 Tessitura model of the β coefficient

The tessitura model for the β coefficient introduced in [5] is a function of the MIDI value m representing its evolution for the whole keyboard of a piano. It can be represented as the sum of two linear asymptotes in the logarithmic scale, whereas these two asymptotes are being described as Treble (b_T) and Bass bridge (b_B) and are characterized as linear functions, parametrized by its slope and constant value, such that the model $\beta_\phi(m)$ can be described as:

$$\beta_\phi(m) = e^{b_B(m)} + e^{b_T(m)} \quad (6)$$

$$= e^{(\phi_1 m+\phi_2)} + e^{(\phi_3 m+\phi_4)} \quad (7)$$

with ϕ being a vector of four elements containing the slope and constant parameters of the linear functions b_B and b_T respectively. All algorithms apart from ours estimate 3 coefficients, denoted $\hat{\beta}$, for each input sound file according to the 3 signal frames which are being used by our algorithm to estimate a single value. A curve fitting is done in a least-squares sense by minimizing the variance of the model $\beta_\phi(m)$ according to (8) with M^* representing the estimates of a single algorithm for one data set. We are using the logarithm of β as well as $\hat{\beta}$ for the objective function to account for the logarithmic behavior of the β coefficient.

$$O_2 = \frac{1}{2}\sum_{m}^{M^*} |\log(\hat{\beta}(m)) - \log(\beta_\phi(m))|^2 \quad (8)$$

Again we are using the scaled Conjugate Gradient method [10] to obtain the tessitura model $\beta_\phi(m)$ with minimum variance using the gradients (9) and (10) for optimizing the parameters for the functions b_B and b_T with i either being set to 1 or 3 for eq. (9) or set to 2 or 4 for eq. (10). The four initial values for vector ϕ are choosen as $[-0.09, -6.87, 0.09, -13.70]^T$.

$$\frac{\partial O_2}{\partial \phi_{1|3}} = \sum_{m}^{M^*} |\log(\hat{\beta}(m)) - \log(\beta_\phi(m))| \frac{m e^{(\phi_i m+\phi_{(i+1)})}}{\beta_\phi(m)} \quad (9)$$

$$\frac{\partial O_2}{\partial \phi_{2|4}} = \sum_{m}^{M^*} |\log(\hat{\beta}(m)) - \log(\beta_\phi(m))| \frac{e^{(\phi_{(i-1)} m+\phi_{(i)})}}{\beta_\phi(m)} \quad (10)$$

As the estimation algorithms may give highly noisy results especially for the upper pitch range we delimit the usage of $\hat{\beta}$ values to a range which is logarithmically close to the initial value by accepting only values which are smaller than ten times the initial function value and bigger than one tenth of it. This is demonstrated in fig. 3, but to finally compute the variance $\sigma^2 = 2N^{-1}O_2$ we take all N estimations of $\hat{\beta}$ into account. The variance according

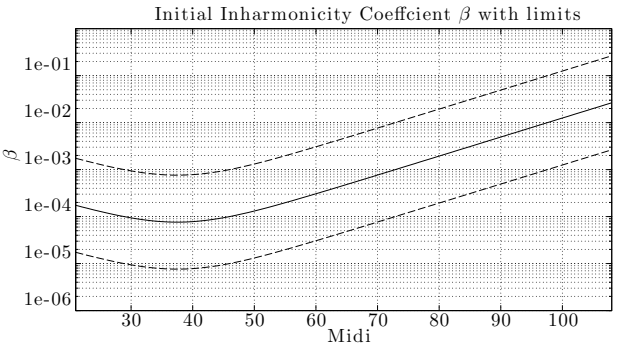

Figure 3. The initial model $\beta_\phi(m)$ (solid) and limits (dashed) for adaptation

to all estimations of $\hat{\beta}$ of one algorithm on data set can be used to determine its estimation accuracy, because we can assume the inharmonicity coefficient of one piano to roughly follow our tessitura model for β. We can further state, that the instruments original β coefficient is equal

for all recordings of the same note of this instrument and constant along time. Therefore, each instrument exhibits a certain variance due to slight tuning errors of its inharmonicity. This variance is unknown and reflects the lower boundary for every estimation algorithm. As all our algorithms estimate either a single inharmonicity value per frame of each sound sample (MAT, ICF, NMF) or a single value per sound sample (CG), the more these values are varying, the less accurate this algorithm has to be. Therefore, we can use the overall variance of the inharmonicity estimations of one algorithm for one data set to determine its accuracy performance.

4.2 Evaluation on artificial data

The sounds have been generated by additive synthesis using eq. (1) to generate the partials frequencies with the β coefficients taken from the initial tessitura model $\beta_\phi(m)$ for each corresponding fundamental frequency, a decaying spectral envelope as well as a simple Attack–Release temporal envelope. The sounds do not include any kind of noise.

We estimated the β values with all methods for all synthesized sounds and measured their deviations from the original values used for synthesis. Fig. 4 shows the resulting relative errors as percentage of the original β value denoted $\bar{\beta}$. As can be seen in fig. 4 the MAT, NMF and CG methods outperform the ICF method with relative errors below 0.1% until MIDI index 86 (D6). Above that index, only the NMF and CG method stay below 0.1% or even drop further down. The estimated tessitura models of all algorithms for the artificial set are shown in fig. 5 and their resulting overall variance of the estimated $\hat{\beta}$ is depicted in fig. 6. The extremely high variance of the results for the MAT and ICF is especially caused by the low estimation accuracy for high pitches (MIDI index values above 85). The increased variance of the NMF method is due to estimation errors around MIDI index 35 at which the inharmonicity coefficient reaches its absolute minimum. Hence, our proposed CG outperforms the MAT and ICF methods significantly in terms of overall variance as it almost never shows an accuracy error of more than 0.1%.

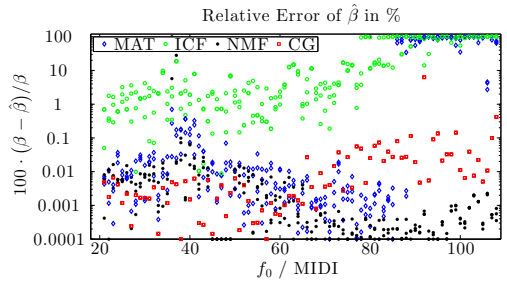

Figure 4. Error in estimation of β given as percentage.

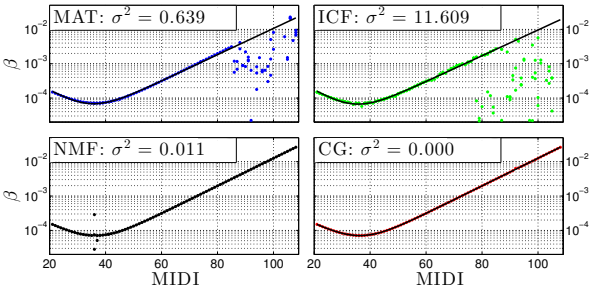

Figure 5. Estimated $\hat{\beta}$ for the artificial data set.

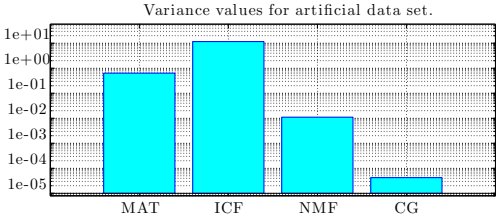

Figure 6. Variance of measurements on artificial data.

4.3 Evaluation on recorded data

The RWC piano library contains recordings of 3 different grand pianos. Each piano has been recorded for all pitches in 3 different intensity levels (*pp*, *mf* and *ff*). The piano set of the IRCAM Solo instruments library also contains recordings for all pitches but with up to 7 intensity levels per pitch and as it has been recorded with 2 discrete channels, we treat these separately. It can be seen in the figures 7 to 11, that the NMF as well as our proposed CG method show especially in the upper pitch range significantly less noise in the estimation of $\hat{\beta}$ compared to the ICF and MAT methods. This seems to be caused by the adaptive noise level used by the NMF method and the peak classification used by CG for selecting reasonable partials. Also, the use of a Kulback-Leibler-divergence with euclidean distance

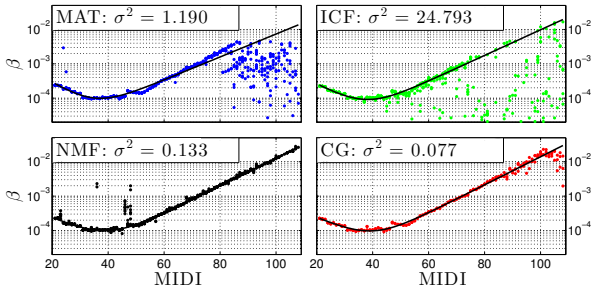

Figure 7. Estimated $\hat{\beta}$ for RWC piano 1

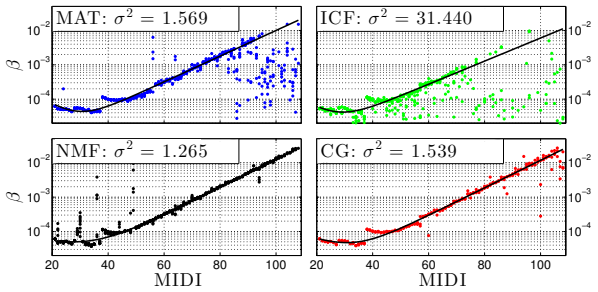

Figure 8. Estimated $\hat{\beta}$ for RWC piano 2

(NMF) and a minimum variance method (CG) for estimating β shows to be clearly superior to a heuristic grid search (ICF) or a median method (MAT). The CG method only shows a slightly higher variance for the RWC 2 data set, whereas it outperforms NMF on all other data sets up to a factor of 20 for the RWC 3 data set . The overall estima-

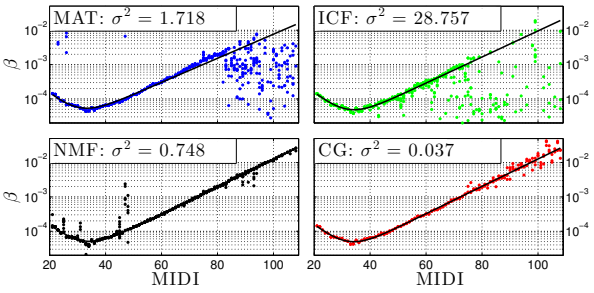

Figure 9. Estimated $\hat{\beta}$ for RWC piano 3

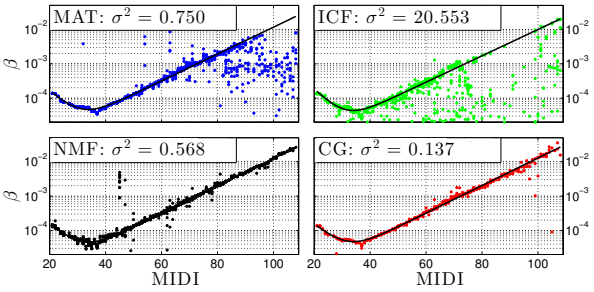

Figure 10. Estimated $\hat{\beta}$ for IRCAM Solo Instrument piano left channel

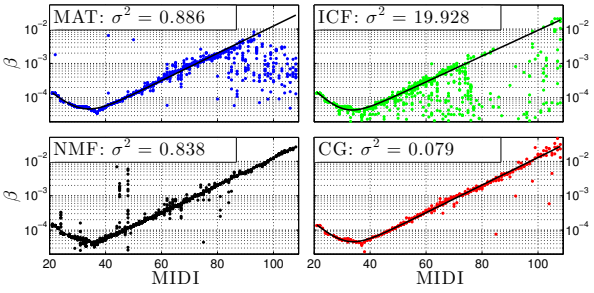

Figure 11. Estimated $\hat{\beta}$ for IRCAM Solo Instrument piano right channel

tion performance is demonstrated in fig. 12. Here, the averaged variance values from all data sets are shown as bars, whereas their minimum and maximum values are given as error bars. It can be observed, that the CG method has the least variance closely followed by the NMF method. The ICF method is far from being accurate, whereas the MAT method rates third. In terms of computational performance, as shown in 13, the MAT method is by far the fastest method, but it clearly lacks in estimation accuracy in the upper pitch range, whereas our proposed method CG outperforms NMF which showed similar estimation results as well as the ICF method.

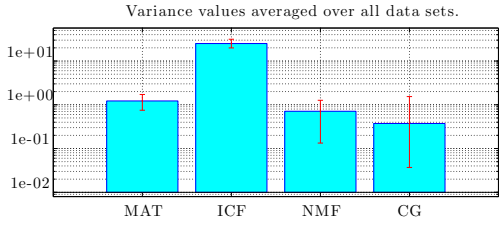

Figure 12. Averaged variance of measurements on real world data according to the tessitura model. The error bars indicate the minimum and maximum variance values among all data sets.

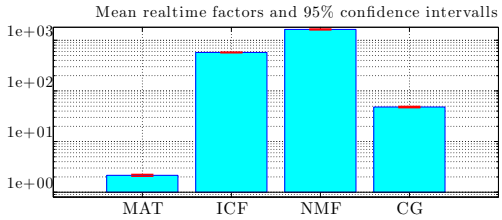

Figure 13. Processing real-time factors for all 4 algorithms averaged for all data sets with 95% confidence intervalls.

5. CONCLUSION

In this paper we gave an overview about three recent approaches (ICF, MAT and NMF) for the estimation of the inharmonicity coefficient and fundamental frequency of inharmonic instrument sounds. We pointed out some issues which are not well addressed in these previous methods and showed possible solutions for these drawbacks with our proposed algorithm. In the evaluation we have shown that for synthetic data with known inharmonicity our proposed algorithm works below an average estimation error in β of 0.1% which clearly outperforms the ICF and MAT method and showed similar accuracy as the NMF method. For real world signals our proposed method again significantly outperforms the MAT and ICF algorithms and showed superior performance in computational efficiency compared with the NMF method which showed a similar estimation accuracy.

Hence, this article shows that a peak selection algorithm with adaptive noise and sidelobe rejection paired with a minimum variance based parameter estimation is a suitable strategy for a robust detection of the inharmonicity coefficient and a signals fundamental frequency.

6. ACKNOWLEDGEMENTS

The authors would like to thank the authors of [3], namely Matthieu Hodgkinson as well as the authors of [4], namely François Rigaud for sharing their sources for the evaluation and their precious help sorting out problems using them. The authors also like to thank the numerous reviewers for their precious remarks and helpful suggestions for improving this document.

This research has been financed as part of the french ANR project Sample Orchestrator 2.

7. REFERENCES

[1] H.Fletcher, E. D. Blackham, and R. Stratton, "Quality of piano tones," *J. Acoust. Soc. Am.*, vol. 34, no. 6, pp. 749 – 761, 1962.

[2] A. Galembo and A. Askenfelt, "Signal representation and estimation of spectral parameters by inharmonic comb filters with application to the piano," *IEEE Transactions On Speech And Audio Processing*, vol. 7, no. 2, pp. 197 – 203, March 1999.

[3] M. Hodgkinson, J. Wiang, J. Timoney, and V. Lazzarini, "Handling inharmonic series with median-adjustive trajectories," in *12th International Conference on Digital Audio Effects (DAFx-09)*, September 2009.

[4] F. Rigaud, B. David, and L. Daudet, "Piano sound analysis using non-negative matrix factorization with inharmonicity constraint," in *Proc. of the 20th European Signal Processing Conference (EUSIPCO 2012)*, August 2012, pp. 2462–2466.

[5] ——, "A parametric model of piano tuning," in *14th International Conference on Digital Audio Effects (DAFx-11)*, September 2011.

[6] J. Rauhala, H.-M. Lehtonen, and V. Valimäki, "Fast automatic inharmonicity estimation algorithm," *J. Acoust. Soc. Am.*, vol. 121, no. 5, pp. EL184 – EL189, May 2007.

[7] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[8] M. Zivanovic, A. Röbel, and X. Rodet, "A new approach to spectral peak classification," in *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, Septembre 2004, pp. 1277–1280. [Online]. Available: http://articles.ircam.fr/textes/Zivanovic04a/

[9] M. Abe and J. Smith, "CQIFFT: Correcting bias in a sinusoidal parameter estimator based on quadratic interpolation of FFT magnitude peaks," Stanford University, Department of Music, Tech. Rep. STAN-M-117, 2004. [Online]. Available: https://ccrma.stanford.edu/STANM/stanms/stanm117/stanm117.pdf

[10] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *NEURAL NETWORKS*, vol. 6, no. 4, pp. 525–533, 1993.

[11] M. Goto and T. Nishimura, "Rwc music database: Music genre database and musical instrument sound database," in *ISMIR*, 2003, pp. 229–230.