# MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION USING GAUSSIAN SMOOTHNESS AND SHORT CONTEXT

**Antonio Pertusa, José M. Iñesta**
University of Alicante, Spain
Departamento de Lenguajes y
Sistemas Informáticos

## ABSTRACT

A multiple fundamental frequency estimator is presented in this work. At each time frame, a set of fundamental frequencies is found in a frame by frame analysis taking into account the spectral smoothness measure described in [1] and the information contained in adjacent frames.

## 1 INTRODUCTION

This work is an extension of the research presented in [1] that was evaluated in MIREX 2007 [2]. In the previous approach, a set of fundamental frequency candidates was selected at each time frame, then all the possible candidate combinations were generated, and the combination with the highest salience was selected. The salience of a combination was computed by taking into account the sum of the harmonic amplitudes and the spectral smoothness of its candidates [1].

Most instruments have spectral patterns that tend to be smooth, and this characteristic is used for estimating the fundamental frequencies that are present in the signal. An interpolation method was also introduced to deal with some harmonic overlap situations.

In [1], each frame was analyzed, yielding a combination of fundamental frequencies that maximized a salience measure. One of the main limitations of this approach is that the window size often used in multiple fundamental frequency estimation (93 ms) is relatively short to detect the fundamental frequencies, even for an expert musician. Context is very important in music to disambiguate certain situations so, in the work presented, short context information is also considered to get a combination of pitches at each frame.

## 2 METHODOLOGY

This new approach is based on [1], but considering information about adjacent frames. Instead of selecting the combination with highest salience at each time frame, short context information is taking into account to get the salience of each combination of pitches, performing a temporal smoothing.

For grouping similar information across time, a set of fundamental frequency combinations are generated at each time frame and the frequencies of each combination $\mathcal{C}$ are converted into MIDI pitches.

For example, the combination $\mathcal{C} = \{261, 416\}$ Hz is converted into $\mathcal{C}' = \{60, 68\}$. In order to have unique combinations in each time frame, if more than one combination with the same pitches is found in a single frame, only the combination with highest salience is kept, removing duplicates with lower saliences.

At a target frame $t$, the new salience $S(\mathcal{C}'(t))$ of a combination $\mathcal{C}'$ is calculated as:

$$S(\mathcal{C}'(t)) = \sum_{i=t-k}^{t+k} S(\mathcal{C}'_i(t)) \quad (1)$$

Therefore, the saliences of the same pitch combinations than those in $\mathcal{C}'$ in the $2k$ adjacent frames are summed to get the salience of the target combination $\mathcal{C}'(t)$. Different values of $k$ were tested, and the best results were obtained with $k = 2$, i.e., considering 2 previous frames, 2 posterior frames and the target frame to get the salience of a combination.

Finally, the maximum salience is selected to get the pitches at the target frame $t$.

$$S(t) = \max_i \{S(\mathcal{C}'_i(t))\} \quad (2)$$

This new approach increases importantly the robustness of the system in the test set used for evaluation, and it allows to remove two parameters added in [1] to avoid local false positives. These parameters are the minimum number of harmonics ($\eta$) to select a spectral peak as a $f_0$ candidate, and the minimum loudness of a peak to be selected as a fundamental frequency candidate.

The equation to get the salience $S$ of a combination in a single frame was also modified respect to [1], increasing the importance to the smoothness value $\sigma(c)$.

$$S = [l(c) \cdot \sigma^4(c)]^2 \quad (3)$$

The harmonic search method was also changed respect to [1]. In the previous work, a constant range $hf_0 \pm f_r$ around each harmonic frequency $hf_0$ for $h = 2, 3, ...$ was

| id | Prec | Rec | Acc | $E_{tot}$ | $E_{subs}$ | $E_{miss}$ | $E_{fa}$ |
|---|---|---|---|---|---|---|---|
| YRC2 | 0.741 | 0.78 | 0.665 | 0.426 | 0.108 | 0.127 | 0.19 |
| YRC1 | 0.698 | 0.741 | 0.619 | 0.477 | 0.129 | 0.129 | 0.218 |
| PI2 | 0.832 | 0.647 | 0.618 | 0.406 | 0.096 | 0.257 | 0.053 |
| RK | 0.698 | 0.719 | 0.613 | 0.464 | 0.151 | 0.13 | 0.183 |
| PI1 | 0.824 | 0.625 | 0.596 | 0.429 | 0.101 | 0.275 | 0.053 |
| VBB | 0.714 | 0.615 | 0.54 | 0.544 | 0.118 | 0.267 | 0.159 |
| DRD | 0.541 | 0.66 | 0.495 | 0.731 | 0.245 | 0.096 | 0.391 |
| CL2 | 0.671 | 0.56 | 0.487 | 0.598 | 0.148 | 0.292 | 0.158 |
| EOS | 0.591 | 0.546 | 0.467 | 0.649 | 0.21 | 0.244 | 0.194 |
| EBD2 | 0.713 | 0.493 | 0.452 | 0.599 | 0.146 | 0.362 | 0.092 |
| EBD1 | 0.674 | 0.498 | 0.447 | 0.629 | 0.161 | 0.341 | 0.127 |
| MG | 0.481 | 0.57 | 0.427 | 0.816 | 0.298 | 0.133 | 0.385 |
| CL1 | 0.358 | 0.763 | 0.358 | 1.68 | 0.236 | 0.001 | 1.443 |
| RFF1 | 0.506 | 0.226 | 0.211 | 0.854 | 0.183 | 0.601 | 0.071 |
| RFF2 | 0.509 | 0.191 | 0.183 | 0.857 | 0.155 | 0.656 | 0.047 |

**Table 1**: Results of multiple $f_0$ estimation task

considered, to allow some harmonic deviations. The closest peak to the center of this margin was set as a harmonic partial and, if no peak was found within this margin, the harmonic was considered as missing. Now, a triangular window centered in $hf_0$, is used to weight the amplitudes of the peaks within this region in order to choose the peak with maximum weighted value.

## 3 POSTPROCESSING

Two different postprocessing techniques were proposed to remove some local errors. Both share the previous methodology.

### 3.1 PI1 method

Sometimes, partials that belongs to one candidate are assigned in a given time frame to other pitch. To avoid temporal discontinuities in the detection, a weighted acyclic directed graph (wDAG) organized by layers is built. Each layer represents a time frame. The nodes of the graph correspond to the $n$ combinations with highest salience at each time frame. The edges of the graph correspond to the inverse of the salience of the destination node multiplied by the loudness differences between two combinations. This loudness difference is computed as the sum of the absolute differences of each note loudness. Finally, the shortest path is found using the Dijkstra algorithm.

### 3.2 PI2 method

This postprocessing technique is described and used in [1]. If a note is shorter than a given minimum duration, it is removed, and if two pitches are separated with a silence shorter than a minimum silence duration, they are glued. These two parameters have changed from the previous work; the minimum note duration (about 20ms), and the minimum silence duration (about 50ms).

## 4 RESULTS

In the multiple $f_0$ estimation task, both systems yielded competitive results and a very good performance. PI2 had a high accuracy, obtaining the highest precision among the systems analyzed, and the lowest error ($E_{tot}$) in the metrics proposed by Polliner and Ellis in [3]. In the Tukey-Kramer HSD significance tests, the first six algorithms ordered by accuracy didn't show significant differences using $p < 0.05$.

The system was presented for the tracking note contours task. Although the algorithm was not designed for this task (for example, no partial tracking is done), the results were satisfactory. In this task, the postprocessing stage PI1 yielded better results than PI2, probably because it favors some temporal continuity in the detection.

| id | Prec | Rec | Avg. F-measure | Avg. Overlap |
|---|---|---|---|---|
| YRC | 0.307 | 0.442 | 0.355 | 0.890 |
| RK | 0.312 | 0.382 | 0.337 | 0.884 |
| ZR3 | 0.256 | 0.314 | 0.278 | 0.874 |
| ZR2 | 0.236 | 0.306 | 0.263 | 0.874 |
| ZR1 | 0.233 | 0.303 | 0.261 | 0.875 |
| PI1 | 0.201 | 0.333 | 0.247 | 0.862 |
| EOS | 0.228 | 0.255 | 0.236 | 0.856 |
| VBB | 0.162 | 0.268 | 0.197 | 0.829 |
| PI2 | 0.145 | 0.301 | 0.192 | 0.854 |
| EBD1 | 0.165 | 0.200 | 0.176 | 0.865 |
| EBD2 | 0.153 | 0.178 | 0.158 | 0.845 |
| RFF2 | 0.037 | 0.030 | 0.032 | 0.645 |
| RFF1 | 0.034 | 0.025 | 0.028 | 0.683 |

**Table 2**: Results of note tracking task

| id | runtime |
|---|---|
| MG | 99 |
| PI2 | 792 |
| PI1 | 955 |
| VBB | 2081 |
| CL1 | 2430 |
| CL2 | 2475 |
| RK | 5058 |
| EOS | 9328 |
| DRD | 14502 |
| EBD1 | 18180 |
| EBD2 | 22270 |
| YRC1 | 57483 |
| YRC2 | 57483 |
| RFF2 | 70041 |
| RFF1 | 73784 |

| id | runtime |
|---|---|
| PI2 | 790 |
| PI1 | 950 |
| ZR3 | 871 |
| ZR1 | 1415 |
| ZR2 | 1415 |
| VBB | 2058 |
| RK | 5044 |
| EOS | 9328 |
| EBD1 | 18180 |
| EBD2 | 22270 |
| YRC | 57483 |
| RFF1 | 73718 |
| RFF2 | 71360 |

**Table 3**: Runtimes of multiple $f_0$ estimation task (left) and tracking note contours task (right).

## 5 ACKNOWLEDGEMENTS

This work is supported by the Spanish PROSEMUS project with code TIN2006-14932-C02 and the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

## 6 REFERENCES


[1] Pertusa, A., Iñesta, J.M. “Multiple Fundamental Frequency estimation using Gaussian smoothness”. *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2008*, Las Vegas, USA, 2008

[2] Pertusa, A., Iñesta, J.M. “Multiple Fundamental Frequency estimation based on spectral pattern loudness and smoothness”. In *MIREX 2007, fundamental frequency estimation and tracking contest*, Vienna, 2007.

[3] Poliner, G.E. and Ellis, D.P.W., “A Discriminative Model for Polyphonic Piano Transcription”. *EURASIP Journal on Advances in Signal Processing*, 2007