City Research Online

City St George's, University of London

Citation: Benetos, E. & Dixon, S. (2010). Multiple-F0 estimation of piano sounds exploiting spectral structure and temporal evolution. Paper presented at the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, 25 Sep 2010, Makuhari, Japan.

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/2770/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Multiple-F0 Estimation of Piano Sounds Exploiting Spectral Structure and Temporal Evolution

Author 1

Affiliation 1
Removed for
double blind review
{author1}@sapa.org

Abstract

This paper proposes a system for multiple fundamental frequency estimation of piano sounds using pitch candidate selection rules which employ spectral structure and temporal evolution. As a time-frequency representation, the Resonator Time-Frequency Image of the input signal is employed, a noise suppression model is used, and a spectral whitening procedure is performed. In addition, a spectral flux-based onset detector is employed in order to select the steady-state region of the produced sound. In the multiple-F0 estimation stage, tuning and inharmonicity parameters are extracted and a pitch salience function is proposed. Pitch presence tests are performed utilizing information from the spectral structure of pitch candidates, aiming to suppress errors occurring at multiples and sub-multiples of the true pitches. A novel feature for the estimation of harmonically related pitches is proposed, based on the common amplitude modulation assumption. Experiments are performed on the MAPS database using 8784 piano samples of classical, jazz, and random chords with polyphony levels between 1 and 6. The proposed system is computationally inexpensive, being able to perform multiple-F0 estimation experiments in real-time. Experimental results indicate that the proposed system outperforms state-of-the-art approaches for the aforementioned task in a statistically significant manner.

Index Terms: multiple-F0 estimation, resonator time-frequency image, common amplitude modulation

1. Introduction

Multiple-F0 estimation in polyphonic music signals refers to the accurate detection of concurrent notes over a short time segment. It is the core problem in the development of automatic transcription systems, which have applications in music information retrieval, interactive computer systems, and automated musicological analysis [1, 9]. While the problem of pitch estimation for monophonic music signals is considered to be solved, the creation of a system able to accurately detect harmonically-related F0s [16] without setting restrictions on the degree of polyphony and the instrument type still remains an open problem. For an overview on state-of-the-art multiple-F0 estimation systems the reader is referred to [4, 9].

There are several approaches for multiple-F0 estimation of music signals related to the current work. In [8], an iterative spectral subtraction method with polyphony inference is proposed, based on the principle that the envelope of harmonic sounds tends to be smooth. A magnitude-warped power spectrum is used as a data representation and a moving average filter is employed for noise suppression. The system is able to handle inharmonicity and experiments were performed on randomly mixed samples from 30 musical instruments compiled from 4 different sources. In [16], a method for jointly evaluating multiple-F0 hypotheses is presented, which employs harmonicity, spectral smoothness, and synchronicity assumptions - the latter is based on the deviation of partials from their temporal centroid. A score function combining the aforementioned criteria is created and its parameters are optimized using an evolutionary algorithm. Experiments were performed with mixtures originating from the same sources as in [8].

A real-time polyphonic transcription system is proposed in [17], which uses a first-order complex resonator filterbank as a time-frequency representation, called the Resonator Time-Frequency Image (RTFI). F0 candidates are selected according to their pitch energy spectrum value and a set of rules is utilized in order to cancel extra estimated pitches. These rules are based on the number of harmonic components detected for each pitch and the spectral irregularity measure, which measures the concentrated energy around possibly overlapped partials from harmonically-related F0s. Finally, a method for multiple-F0 estimation of piano sounds is developed in [5], which models the spectral envelope of pitches using a smooth autoregressive model constrained by the spectral smoothness principle and models the noise using a moving average model. A pitch salience function that is able to handle tuning and inharmonicity is proposed for initial candidate selection and the candidates are refined using a likelihood function which is dependent on the estimated spectral envelope and noise parameters. Experiments were performed on a database called MAPS, which contains real or synthesized recordings of isolated notes, musical or random chords, as well as music pieces, which were produced by several piano types or using different recording conditions. Results, compared with the method in [8], indicate that the proposed system is particularly able to yield good scores when harmonically-related F0s are present.

In this work, a system for multiple-F0 estimation of isolated piano sounds which uses candidate selection and several rule-based refinement steps is proposed. The RTFI is used as a data representation [17], and preprocessing steps for noise suppression, spectral whitening, and onset detection are utilized in order to make the estimation system robust to noise and recording conditions. A pitch salience function that is able to function in the log-frequency domain and utilizes tuning and inharmonicity estimation procedures is proposed and pitch candidates are selected according to their salience value. The set of candidates is refined using rules regarding the harmonic partial sequence of the selected pitches and the temporal evolution of the partials, in order to minimize errors occurring at multiples

and sub-multiples of the actual F0s. For the spectral structure rules, a more robust formulation of the spectral irregularity measure [17] is proposed, taking into account overlapping partials. For the temporal evolution rules, a novel feature based on the common amplitude modulation (CAM) assumption [11] is proposed in order to suppress estimation errors in harmonically-related F0 candidates. Experiments were performed on the MAPS database [5] using over 8000 classical, jazz, and random piano chords, produced by 9 different piano types and recording conditions. Results indicate that the proposed system outperforms the state-of-the-art approaches developed in [5] and [8] for the same experiment.

The remainder of the paper is as follows. In Section 2, the preprocessing steps used in the proposed system are described. The multiple frequency estimation system is detailed in Section 3. In Section 4 the employed dataset is presented, the experimental procedure is described, and results are discussed. Concluding remarks are drawn and future directions are pointed out in Section 5.

2. Preprocessing

In this section, the preprocessing steps employed by the proposed multiple-F0 estimation system are described. These steps can also be seen in a diagram for the proposed system, which is displayed in Figure 1.

2.1. Resonator Time-Frequency Image

Firstly, the overall loudness of the time-domain input signal $x[n]$ is normalized to 70dB level. As a time-frequency representation, the RTFI was used [17]. The RTFI selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis. It can be formulated as:

$$RTFI(t,\omega) = s(t) * I_R(t,\omega) \tag{1}$$

where

$$I_R(t,\omega) = r(\omega)e^{(-r(\omega)+j\omega)t}. \tag{2}$$

$s(t)$ stands for the input signal, $I_R(t,\omega)$ is the impulse response of the first-order complex resonator filter with oscillation frequency ω and $r(\omega)$ is a decay factor which additionally sets the frequency resolution.

For the specific experiments, a RTFI with constant-Q resolution is selected for the time-frequency analysis, due to its suitability for music signal processing techniques, because the inter-harmonic spacing is the same for all pitches. The time interval between two successive frames is set to 40ms, which is typical for multiple-F0 estimation approaches [9]. A sampling rate of 44100Hz is considered for the input samples and the centre frequency difference between two neighbouring filters is set to 10 cents (the number of bins per octave b is set to 120). The frequency range is set from 27.5Hz (A0) to 12.5kHz (which reaches up to the 3rd harmonic of C8). The employed absolute value of the RTFI will be denoted as $X[n,k]$, where n is the time frame and k the frequency bin.

2.2. Spectral Whitening and Noise Suppression

Spectral whitening is employed in order to flatten the dynamic range of the RTFI bins. Here, a modified version of the real-time adaptive whitening method proposed in [14] is applied. Each band is scaled, taking into account the temporal evolution of the signal, while the scaling factor is dependent only on past frame values and the peak scaling value is exponentially decaying. The following iterative algorithm is applied:

$$\begin{aligned} Y[n,k] &= \begin{cases} \max(|X[n,k]|, c, aY[n-1,k]), & n > 0 \\ \max(|X[n,k]|, c), & n = 0 \end{cases} \\ X[n,k] &\leftarrow \frac{X[n,k]}{Y[n,k]} \end{aligned} \tag{3}$$

where a is the peak scaling value and c is a floor parameter.

In addition, a noise suppression approach similar to the one in [10] was employed, due to its computational efficiency. A half-octave span (60 bins) moving median filter is computed for $X[n,k]$, resulting in noise estimate $N[n,k]$. Afterwards, an additional moving median filter $N'[n,k]$ of the same span is applied, but only including the RTFI bins whose amplitude is less than the respective amplitude of $N[n,k]$. This results in making the noise estimate $N'[n,k]$ robust in the presence of spectral peaks that could affect the noise estimate $N[n,k]$.

2.3. Onset Detection

In order to select the steady-state area of the produced note(s), a spectral flux-based onset detection procedure is applied. The *spectral flux* measures the magnitude changes in each frequency bin which indicate the attack parts of new notes [2]. It can be used effectively for onset detection of notes produced by percussive instruments such as the piano, but its performance decreases for the detection of soft onsets [1]. For the RFTI, the spectral flux using the L1 norm can be defined as:

$$SF[n] = \sum_k HW(|X[n,k]| - |X[n-1,k]|) \tag{4}$$

where $HW(x) = \frac{x+|x|}{2}$ is a half-wave rectifier. The resulting onset strength signal is smoothed using a median filter with a 3 sample span (120ms length), in order to remove spurious peaks. Onsets are subsequently selected from $SF[n]$ by a selection of local maxima, with a minimum peak distance of 120ms. Afterwards, the frames located between 100-300ms after the onset are selected as the steady-state region of the signal and are averaged over time, in order to produce a robust spectral representation of the produced notes.

3. Proposed System

The algorithm that was created for multiple-F0 estimation experiments is described in this section. A diagram showing the stages of the proposed system is displayed in Figure 1.

3.1. Salience Function Generation

In the linear frequency domain, considering a pitch p of a piano sound with fundamental frequency f_{0p} and inharmonicity coefficient β_p, partials are located at frequencies:

$$f_{hp} = hf_{0p}\sqrt{1+(h^2-1)\beta_p} \tag{5}$$

where $h \geq 1$ is the partial index [9, 13]. Consequently in the log-frequency domain, considering a pitch p at bin k_{0p}, overtones are located at bins:

$$k_{hp} = k_{0p} + \left\lceil b \cdot \log_2(h) + \frac{b}{2}\log_2\left(1+(h^2-1)\beta_p\right)\right\rfloor \tag{6}$$

where $b = 120$ refers to the number of bins per octave.

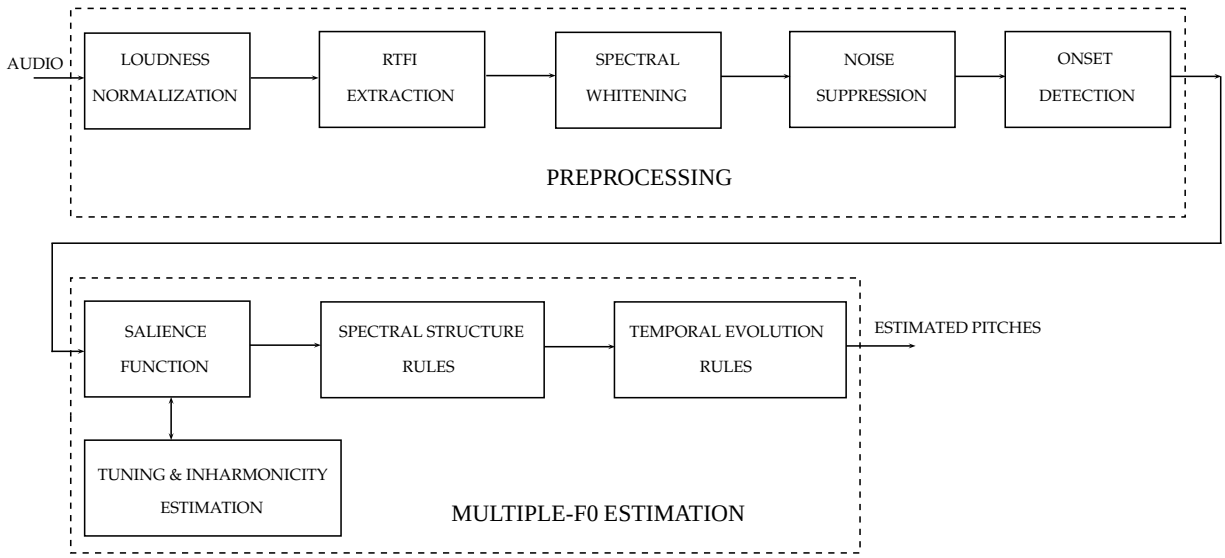

Figure 1: Diagram for the proposed multiple fundamental frequency estimation system.

A pitch salience function $s[p, d_p, \beta_p]$ operating in the log-frequency domain is proposed, which indicates the strength of pitch candidates:

$$s[p, d_p, \beta_p] = \sum_{h=1}^{H} \max_{m_h} \Big\{ Z[k_{hp} + d_p, m_h] \Big\} \qquad (7)$$

where

$$Z[k, m_h] = \sqrt{X\left[k + \left\lceil bm_h + \frac{b}{2}\log_2(1 + (h^2 - 1)\beta)\right\rceil\right]} \qquad (8)$$

and m_h specifies a search range around overtone positions, belonging to the interval (m_h^l, m_h^u), where $m_h^l = \lceil \frac{\log_2(h-1)+(M-1)\log_2(h)}{M} \rfloor$, $m_h^u = \lceil \frac{(M-1)\log_2(h)+\log_2(h+1)}{M} \rfloor$. M is a factor controlling the width of the interval, which for the current experiments was set to 60. The salience function is applied to the averaged steady-state representation shown in Section 2.3.

While the employed salience functions in the linear frequency domain (ie. [10]) used a constant search space for each overtone, the proposed log-frequency salience function sets the search space to be inversely proportional to the partial index. The number of considered overtones H is set to 11 at maximum. Tuning is also considered [15], with a tuning deviation $d_p \in [-4, \ldots, 4]$ for each pitch (thus having a tuning search space of 80 cents around the ideal tuning frequency). The range of the inharmonicity coefficient β_p is set between 0 and $5 \cdot 10^{-4}$, which is typical for piano notes [13].

In order to accurately estimate the tuning factor and the inharmonicity coefficient for each pitch, a two-dimensional maximization procedure using exhaustive search is applied to $s[p, d_p, \beta_p]$ for each pitch $p \in [21, \ldots, 108]$ in the MIDI scale with $k_{0p} = 10(p - 21) + 1$ (corresponding to a note range of A0-C8). This results in a pitch salience function estimate $s'[p]$, a tuning deviation vector and an inharmonicity coefficient vector. Using the information extracted from the tuning and inharmonicity estimation,, a harmonic partial sequence $V[p, h]$ for each candidate pitch and its harmonics (which contains the RTFI values at certain bin) is also stored for further processing.

3.2. Spectral Structure Rules

A set of rules examining the harmonic partial sequence structure of each pitch candidate is applied, which is inspired by work from [1, 17]. These rules aim to suppress peaks in the salience function that occur at multiples and sub-multiples of the actual fundamental frequencies. In the semitone space, these peaks occur at $\pm\{12, 19, 24, 28, \ldots\}$ semitones from the actual pitch.

A first rule for suppressing salience function peaks is setting a minimum number for partial detection in $V[p, h]$, similar to [1, 17]. If $p < 47$, at least three partials out of the first six need to be present in the harmonic partial sequence (since there may be a missing fundamental). If $p \geq 47$, at least four partials out of the first six should be detected. A second rule concerns the salience value, which expresses the sum of the square root of the partial sequence amplitudes. If the salience value is below a minimum threshold (set to 0.2 using the development set explained in Section 4.1), this peak is suppressed. Another processing step in order to reduce processing time is the reduction of the number of pitch candidates [5], by selecting only the pitches with the greater salience values. In the current experiments, 10 candidate pitches are selected from $s'[p]$.

Spectral flatness is another descriptor that can be used for the elimination of errors occurring in subharmonic positions [5]. In the proposed system, the flatness of the first 6 partials of a harmonic sequence is used:

$$Fl[p] = \frac{\sqrt[6]{\prod_{h=1}^{6} V[p, h]}}{\frac{\sum_{h=1}^{6} V[p,h]}{6}} \qquad (9)$$

The ratio of the geometric mean of V to its arithmetic mean gives a measure of smoothness; a high value of $Fl[p]$ indicates a smooth partial sequence, while a lower value indicates fluctuations in the partial values, which could indicate the presence of a falsely detected pitch occurring in a sub-harmonic position. For the current experiments, the lower $Fl[p]$ threshold for suppressing pitch candidates was set to 0.1 after experimentation using the development set, as described in subsection 4.1.

In order to suppress candidate pitches occurring at multiples of the true fundamental frequency, a modified version of

the *spectral irregularity* measure formulated in [17] is proposed. Considering a pitch candidate with fundamental frequency f_0 and another candidate with fundamental frequency lf_0, $l > 1$, spectral irregularity is defined as:

$$SI[p,l] = \sum_{h=1}^{3} V[p,hl] - \frac{V[p,hl-1] + V[p,hl+1]}{2} \quad (10)$$

The spectral irregularity is tested on pairs of harmonically-related candidate F0s (where $f_1 = lf_0$). A high value of $SI[p,l]$ indicates the presence of the higher pitch with fundamental frequency lf_0, which is attributed to the higher energy of the shared partials between the two pitches compared to the energy of the neighbouring partials of f_0.

In this work, the SI is modified in order to make it more robust against overlapping partials that are caused by non-harmonically related F0s [16]. Given the current set of candidate pitches from $s'[p]$, the overlapping partials from non-harmonically related F0s are detected as in [16] and smoothed according to the *spectral smoothness* assumption, which states that the spectral envelope of harmonic sounds should form a smooth contour [8]. For each overlapping partial $V[p,h]$, an interpolated value $V_{interp}[p,h]$ is estimated by performing linear interpolation using its neighbouring partials. Afterwards, the smoothed partial amplitude $V'[p,h]$ is given by $\min(V[p,h], V_{interp}[p,h])$, as in [8]. The proposed spectral irregularity measure, which now takes the form of a ratio for in order to take into account the decreasing amplitude of higher partials, is thus formed as:

$$SI'[p,l] = \sum_{h=1}^{3} \frac{2 \cdot V'[p,hl]}{V'[p,hl-1] + V'[p,hl+1]} \quad (11)$$

For each pair of harmonically-related F0s (candidate pitches that have a pitch distance of $\pm\{12, 19, 24, 28, \ldots\}$) that are present in $s'[p]$, the existence of the higher pitch is determined by the value of SI' (for the current experiments, a threshold of 1.2 was set using the development set).

3.3. Temporal Evolution Rules

Although the SI and the spectral smoothness assumption are able to suppress some harmonic errors, additional information needs to be exploited in order to produce more accurate estimates in the case of harmonically-related F0s. In [16], temporal information was employed for multiple-F0 estimation using the synchoronicity criterion as a part of the F0 hypothesis score function. There, it is stated that the temporal centroid for a harmonic partial sequence should be the same for all partials. Thus, partials deviating from their global temporal centroid indicates an invalid F0 hypothesis. Here, we use the *common amplitude modulation* (CAM) assumption [6, 11] in order to test the presence of a higher pitch in the case of harmonically-related F0s. CAM assumes that the partial amplitudes of a harmonic source are correlated over time and has been used in the past for note separation given a ground truth of F0 estimates [11]. Thus, the presence of an additional source that overlaps certain partials (eg. in the case of an octave where even partials are overlapped) causes the correlation between non-overlapped partials and the overlapped partials to decrease.

To that end, tests are performed for each harmonically-related F0 pair that is still present in $s'[p]$, comparing partials that are not overlapped by any non-harmonically related F0 candidate with the partial of the fundamental. The correlation coefficient is formed as:

$$Corr[p,h,l] = \frac{Cov(X[n,k_{p,1}], X[n,k_{p,hl}])}{\sqrt{Cov(X[n,k_{p,1}])Cov(X[n,k_{p,hl}])}} \quad (12)$$

where $k_{p,h}$ indicates the frequency bin corresponding to the h-th harmonic of pitch p, n denotes the RTFI frame number, l the harmonic relation (eg. for octaves $l = 2$), and $Cov(\cdot)$ stands for the covariance measure. Tests are being taken for each pitch p and harmonics hl, using the same steady-state area used in subsection 2.3 as a frame range. If there is at least one harmonic where the correlation coefficient for a pitch is lower than a given value (in the experiments it was set to 0.8), then the hypothesis for the higher pitch presence is satisfied. In order to demonstrate the various refinement steps used in the salience function, Figure 2 shows the three basic stages of the multiple-F0 estimation system for a synthesized E♭4-G4-B♭4-C5-D5 piano chord.

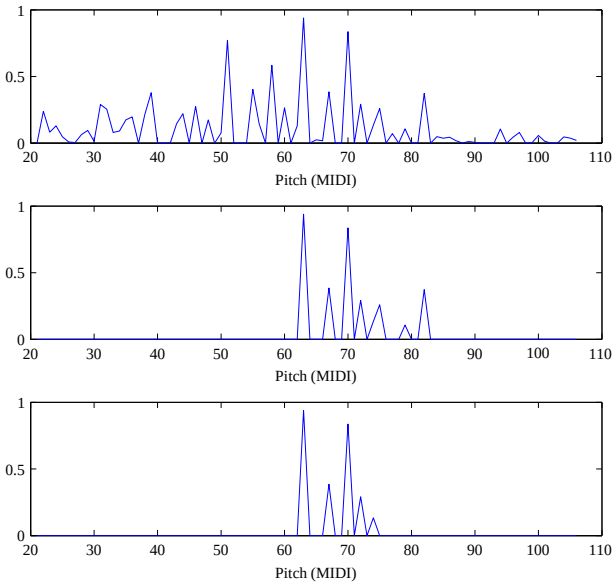

Figure 2: Salience function stages for an E♭4-G4-B♭4-C5-D5 piano chord. From top to bottom, the figures represent (i) The raw salience function (ii) The salience function after the spectral structure rules have been applied (iii) The salience function after the temporal evolution tests have been applied.

4. Evaluation

4.1. Dataset

The proposed multiple-F0 estimation system was tested on the MIDI Aligned Piano Sounds (MAPS) database [5]. It contains real and synthesized recordings of isolated notes, musical chords, random chords, and music pieces, produced by 9 real and synthesized pianos in different recording conditions, containing around 10000 sounds in total. Recordings are stereo, sampled at 44100Hz, while MIDI files are provided as ground truth. For the current experiments, classic, jazz, and randomly generated chords (without any note progression) of polyphony levels between 1 and 6 were employed, while the note range was C2-B6, in order to match the experiments performed in [5]. Each recording lasts about 4 seconds. A development set using

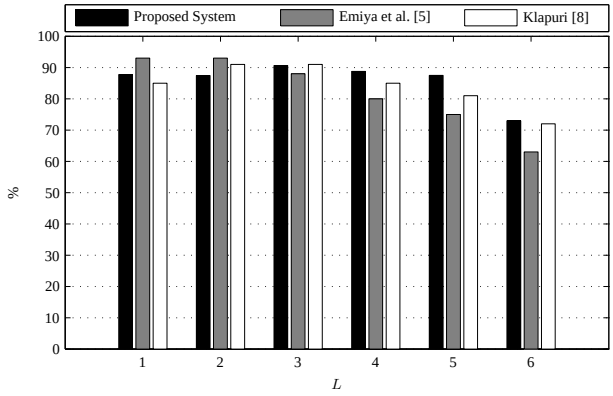

Figure 3: Multiple-F0 estimation results (in F-measure) with unknown polyphony, organized according to the ground truth polyphony level L.

2 pianos (consisting of 1952 samples) is selected while the other 7 pianos (consisting of 6832 samples) are used as a test set.

4.2. Figures of Merit

In order to evaluate the results of the proposed multiple-F0 estimation system, the recall, precision, and F-measure are used:

$$P = \frac{tp}{tp + fp}, \quad R = \frac{tp}{tp + fn}, \quad F = \frac{2PR}{P + R} \qquad (13)$$

where tp is the number of correctly estimated pitches, fp is the number of false pitch detections, and fn is the number of missed pitches. A set or P, R, F is generated for each recording. By varying the system parameters, precision/recall (P/R) curves can be created by placing R values on the x-axis and P values on the y-axis.

4.3. Results

The performance of the proposed multiple-F0 estimation system compared with the results shown in [5] is shown in Figure 3, organized according to the polyphony level of the ground truth (experiments were performed with unknown polyphony). The mean F-measures for polyphony levels $L = 1, \ldots, 6$ are 87.84%, 87.44%, 90.62%, 88.76%, 87.52%, and 72.96% respectively. It should be noted that the subset of polyphony level 6 consists only of 350 samples of random notes and not of classical and jazz chords. As far as precision is concerned, reported rates are high for polyphony levels 2-6, ranging from 91.11% to 95.83%. The lowest precision rate is 84.25% for $L = 1$, where some overtones were erroneously considered as pitches. Recall displays the opposite performance, reaching 96.42% for one-note polyphony, and decreasing with the polyphony level, reaching 87.31%, 88.46%, 85.45%, and 82.35%, and 62.11% for levels 2-6.

Comparing the results with the system in [5] (where the reported F-measures for the same polyphony levels were 93%, 93%, 88%, 80%, 75%, and 63%), it can be seen that the proposed system yields improved results for polyphony levels 3-6, while falling back in the one- and two-note polyphony case. The best improvement is reported for $L = 5$, which is about 12.5%. The algorithm in [5] follows the same pattern when P and R are concerned, reporting high P rates for all polyphony levels and decreasing R rates as polyphony increases. Additional experiments were performed in [5] using the iterative spectral subtraction algorithm proposed by Klapuri in [8], which reached F-measures of about 85%, 91%, 91%, 85%, 81%, and 72% for $L = 1, \ldots, 6$, respectively. In this case, the proposed system performs better for $L = 1, 4, 5, 6$, reporting the best improvement (6.5%) for the 5-note polyphony case, while the worst performance difference is about 3.5% for $L = 2$.

In terms of a general comparison between the 3 systems, a weighted F-measure was used, weighting the various F for polyphony levels 1-6 with their respective set size, since the global F-measure was not reported in [5]. For the proposed system, the actual global F is 87.48%. For the algorithm in [5], the estimated global F is 83.70%, while for the algorithm of [8] used in [5], it is 85.25%.

Concerning the statistical significance of the proposed method's performance compared to the methods in [5, 8], the recognizer comparison technique described in [7] was employed. The number of pitch estimation errors of the two methods is assumed to be distributed according to the binomial law. The error rate of the proposed method is $\hat{p}_1 = 0.1252$, while the average error rate of the two methods in [5] is $\hat{p}_2 = 0.1630$ and $\hat{p}_3 = 0.1475$. Taking into account that the test set size $S = 6832$ and considering 95% confidence ($\alpha = 0.05$), it can be seen that $\hat{p}_2 - \hat{p}_1 \geq z_\alpha \sqrt{2\hat{p}/S}$, where z_α can be determined from tables of the Normal law ($z_{0.05} = 1.65$) and $\hat{p} = \frac{\hat{p}_1 + \hat{p}_2}{2}$. Likewise, it can be seen that $\hat{p}_3 - \hat{p}_1 \geq z_\alpha \sqrt{2\hat{p}/S}$, where in this time $\hat{p} = \frac{\hat{p}_1 + \hat{p}_3}{2}$. This indicates that the performance of the proposed multiple-F0 method is significantly better when compared with the methods in [5, 8].

Another issue for comparison is the matter of computational complexity, where the algorithm in [5] being reported to require a process time of about 150×real time, while the proposed system is able to estimate pitches faster than real time (implemented in Matlab), with the bottleneck being the RTFI computation; all other processes are almost negligible regarding computation time. This makes the proposed approach attractive as a potential application for automatic polyphonic music transcription.

In [5], additional results are reported using a subset of 97 recordings which only contains octaves. The system in [5] yielded an F-measure of 81%, while the algorithm in [8] reached 77%. Here, the reported mean F_{oct} is 84.59%, with $P_{oct} = 90.59\%$ and $R_{oct} = 84.12\%$. The improved performance of the proposed system on octave detection could be attributed to the octave presence tests that were performed using the SI measure as well as on the temporal evolution tests using the partial correlation. In contrast, the method in [5] uses the smoothness of the partial envelope as a pitch presence indication, which is not sufficient for detecting octaves. Additional insight to the performance of the octave detection experiments is given in the form of a P/R curve with varying SI' in Figure 4. When $SI' = 0.25$, F_{oct} reaches a value of 87.59%, while when using the SI' threshold for the whole system the F_{oct} drops about 4%. When the value of SI' reaches 5, the recall drops to 50%, which indicates that only the lower pitches of the octaves are selected.

5. Conclusions

In this work, a system for multiple fundamental frequency estimation of piano sounds was proposed. The constant-Q resonator time-frequency image was selected as a mid-level data representation, while techniques for noise suppression, spec-

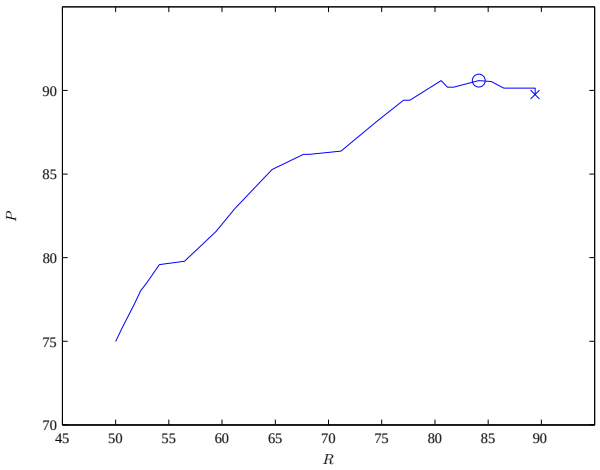

Figure 4: P/R curve for the octave detection experiments with $SI' \in \{0, 5\}$. The circle marker corresponds to the selected SI' for the whole system (with $F_{oct} = 84.59\%$) and the cross marker corresponds to the optimal SI' value for the octave experiments only ($F_{oct} = 87.59\%$).

tral whitening, and onset detection were employed in order to make the subsequent analysis robust. A log-frequency salience function was proposed, being able to handle tuning and inharmonicity estimation, while pitch candidates were selected and refined according to a set of rules based on spectral characteristics. A novel procedure for suppressing errors by harmonically-related F0s was proposed, using the common amplitude modulation assumption, which takes the form of partial correlation tests. Experiments were performed on a large dataset of piano recordings containing samples that were created using different sources and recording conditions. The system reports increased pitch estimation performance when compared to state-of-the-art approaches. Statistical significance tests were carried out in order to verify the proposed method's superiority. In addition, the system was able to address the octave detection problem by employing tests on harmonically-related F0s.

In the future, experiments will be performed on datasets consisting of various instrument types, such as the 30 instrument random mixtures dataset used in [8]. To that end, automatic adaptation of the proposed system parameters according to the spectral envelope shape and the partial evolution of the produced notes is necessary. It should be noted that, although the method can be extended in order to cover several instrument types, the detection of notes produced by extremely inharmonic instruments such as marimba or vibraphone cannot be supported by the current system. In addition, a robust onset detection algorithm will be developed in order to accurately detect soft onsets produced by pitched non-percussive instruments. More emphasis will also be given to the matter of correctly estimating the pitch of complex combinations of harmonically-related notes, which still remains an open problem in the literature. Finally, the multiple-F0 estimation algorithm will be incorporated into an automated music transcription system.

6. References

[1] J. P. Bello, "Towards the automated analysis of simple polyphonic music: a knowledge-based approach," *PhD Diss.*, Queen Mary, University of London, Jan. 2003.

[2] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection of music signals," *IEEE Trans. Speech and Audio Proc.*, Vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[3] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoustical Society of America*, Vol. 89, no. 1, pp. 425–434, Jan. 1991.

[4] A. de Cheveigné, "Multiple F0 estimation," in D. L. Wang and G. J. Brown (Eds), *Computational Auditory Scene Analysis, Algorithms and Applications*, IEEE Press/Wiley, pp. 45–79, 2006.

[5] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech and Language Processing*, Vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[6] D. Gunawan and D. Sen, "Identification of partials in polyphonic mixtures based on temporal envelope similarity," in Proc. *AES 123rd Convention*, Oct. 2007.

[7] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error estimates?," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, no. 1, pp. 52–64, Jan. 1998.

[8] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 11, no. 6, pp. 804–816, Nov. 2003.

[9] A. Klapuri and M. Davy (Eds), *Signal Processing Methods for Music Transcription*, Springer-Verlag, New York, 2006.

[10] A. Klapuri, "A method for visualizing the pitch content of polyphonic music signals," in Proc. *10th Int. Conf. Music Information Retrieval*, pp. 615–620, Oct. 2009.

[11] Y. Li, J. Woodruff, and D. L. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 17, no. 7, pp. 1361–1371, Sep. 2009.

[12] Music Information Retrieval Evaluation eXchange (MIREX), `http://music-ir.org/mirexwiki/`.

[13] L. I. Ortiz-Berenguer, F. J. Casajús-Quirós, M. Torres-Guijarro, and J. A. Beracoechea, "Piano transcription using pattern recognition: aspects on parameter extraction," in Proc. *7th Int. Conf. Digital Audio Effects*, pp. 212–216, Oct. 2004.

[14] D. Stowell and M. Plumbley, "Adaptive whitening for improved real-time audio onset detection," in Proc *Int. Computer Music Conf.*, pp. 312–319, Aug. 2007.

[15] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in Proc. *2008 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 109–112, Apr. 2008.

[16] C. Yeh, "Multiple fundamental frequency estimation of polyphonic recordings," *PhD Diss.*, École Doctorale Edite, Université Paris VI - Pierre et Marie Curie, Jun. 2008.

[17] R. Zhou, "Feature Extraction of Musical Content for Automatic Music Transcription," *PhD Diss.*, Swiss Federal Institute of Technology, Lausanne, Oct. 2006.