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

Emmanouil Benetos and Simon Dixon
Centre for Digital Music, Queen Mary University of London, Mile End Road, London E1 4NS, UK
{emmanouil.benetos, simon.dixon}@elec.qmul.ac.uk

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:

\[
RTFI(t, \omega) = s(t) * IR(t, \omega)
\]

\[
IR(t, \omega) = r(\omega)e^{-r(\omega)+j\omega}t.
\]

where \( s(t) \) stands for the input signal, \( IR(t, \omega) \) is the impulse response of the first-order complex resonator filter with oscillation frequency \( \omega \) 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:

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

\[
X[n, k] \leftarrow X[n, k] \frac{Y[n, k]}{Y[n, k]}
\]

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 RTFI, the spectral flux using the L1 norm can be defined as:

\[
SF[n] = \sum_k HW(|X[n, k]| - |X[n - 1, k]|)
\]

where \( HW(x) = \frac{\sin(x/2)}{x/2} \) 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 \( \beta_p \), partials are located at frequencies:

\[
f_{kp} = hf_0p \sqrt{1 + (h^2 - 1)\beta_p}
\]

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:

\[
k_{hp} = k_{0p} + \left[ b \cdot \log_2(h) + \frac{b}{2} \log_2 \left( 1 + (h^2 - 1)\beta_p \right) \right]
\]

where \( b = 120 \) refers to the number of bins per octave.
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:

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

where

\[
Z[k, m_h] = \sqrt{X[k + \left( b m_h + \frac{b}{2} \log_2(1 + (h^2 - 1)\beta) \right]}
\]  \hspace{1cm} (8)

and \(m_h\) specifies a search range around overtone positions, belonging to the interval \((m^1_h, m^2_h)\), where \(m^1_h = \frac{\log_2(h-1) + (M-1)\log_2(h)}{b\log_2(1 + h^2 - 1)\beta} - 1\), \(m^2_h = \frac{(M-1)\log_2(h) + \log_2(h^2 + 1)}{b\log_2(1 + h^2 - 1)\beta}\). \(M\) 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 \(\beta_p\) is set between 0 and 5\(\times\)10\(^{-3}\), 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 \(k_{hp} = 10(p - 21) + 1\) (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:

\[
F_l[p] = \sqrt[6]{\prod_{h=1}^{6} V[p, h]} / \sqrt{\sum_{h=1}^{6} V[p, h]}
\]  \hspace{1cm} (9)

The ratio of the geometric mean of \(V\) to its arithmetic mean gives a measure of smoothness; a high value of \(F_l[p]\) 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 \(F_l[p]\) 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 \( l f_0 \), \( l > 1 \), spectral irregularity is defined as:

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

(10)

The spectral irregularity is tested on pairs of harmonically-related candidate F0s (where \( f_2 = l f_1 \)). A high value of \( SI[p, l] \) indicates the presence of the higher pitch with fundamental frequency \( l f_0 \), 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:

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

(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 synchronicinity 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 candi-
The mean F-measures for polyphony levels
truth (experiments were performed with unknown polyphony).
comparied with the results shown in [5] is shown in Figure 4.3. Results
on the y-axis.

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 estima-
tion system, the recall, precision, and F-measure are used:

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

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 sys-
tem 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% re-
spectively. It should be noted that the subset of polyphony level
6 consists only of 350 samples of random notes and not of clas-
sical 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 re-
ported 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 experi-
ments were performed in [5] using the iterative spectral sub-
traction 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 improve-
ment (6.5%) for the 5-note polyphony case, while the worst per-
formance 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 em-
ployed. The number of pitch estimation errors of the two meth-
ods is assumed to be distributed according to the binomial law.
The error rate of the proposed method is \( p_1 = 0.1252 \), while
the average error rate of the two methods in [5] is \( p_2 = 0.1630 \)
and \( p_3 = 0.1475 \). Taking into account that the test set size
\( S = 6832 \) and considering 95% confidence (\( \alpha = 0.05 \)), it can
be seen that \( p_2 - p_1 \geq z \alpha \sqrt{2pS}/S \), where \( z \alpha \) can be determined
from tables of the Normal law (\( z \alpha = 1.65 \)) and \( \hat{p} = \frac{\hat{p}_1 + \hat{p}_2}{2} \).
Likewise, it can be seen that \( p_3 - p_1 \geq z \alpha \sqrt{2pS}/S \), where in
this time \( \hat{p} = \frac{\hat{p}_3 + \hat{p}_1}{2} \). 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 computa-
tional complexity, where the algorithm in [5] being reported
to require a process time of about 150 ms 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 regard-
ing computation time. This makes the proposed approach attractive
as a potential application for automatic polyphonic music tran-
scription.

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 \( F_{oct} = 84.59\% \), with
\( P_{oct} = 90.59\% \) and \( R_{oct} = 84.12\% \). The improved perfor-
mance of the proposed system on octave detection could be at-
tributed 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 indica-
tion, 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 es-
timation of piano sounds was proposed. The constant-Q res-
onator time-frequency image was selected as a mid-level data represen-
tation, while techniques for noise suppression, spec-
6. Acknowledgement

The authors would like to thank Valentin Emiya for generously providing the MAPS dataset. This work was supported by a Westfield Trust Research Studentship (Queen Mary, University of London).

7. References


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 $F_{\text{oct}} = 84.59\%$) and the cross marker corresponds to the optimal $SI'$ value for the octave experiments only ($F_{\text{oct}} = 87.59\%$).