MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION USING SPECTRAL STRUCTURE AND TEMPORAL EVOLUTION RULES

Emmanouil Benetos and Simon Dixon
Centre for Digital Music, Queen Mary University of London
{emmanouil.benetos,simon.dixon}@elec.qmul.ac.uk

ABSTRACT

This paper describes a method submitted for the MIREX 2010 Multiple Fundamental Frequency Estimation & Tracking Task 1, which uses pitch candidate selection rules employing spectral structure and temporal evolution. For preprocessing, the Resonator Time-Frequency Image of the input signal is employed as a time-frequency representation, a noise suppression model is used, and a spectral whitening procedure is performed. Also, tuning and inharmonicity parameters are extracted for the complete recording and a frame-by-frame pitch salience function is generated. 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. Additional tests for the estimation of harmonically related F0s are performed over time, based on the common amplitude modulation assumption.

1. INTRODUCTION

Automatic music transcription is the process of converting an audio recording into a symbolic representation using musical notation. The core problem in automatic transcription is the estimation of concurrent pitches in a time frame, also called multiple-F0 estimation. Important subtasks for automatic music transcription also include onset/offset detection, loudness estimation, instrument recognition, and extraction of rhythmic information.

The proposed system which is submitted for the MIREX Multiple Fundamental Frequency Estimation & Tracking Task 1 (frame-by-frame evaluation) offers a computationally inexpensive way for multi-pitch estimation, using candidate selection and several rule-based refinement steps. A diagram showing the system stages is displayed in Figure 1.

2. PREPROCESSING

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 resonator time-frequency image (RTFI) was used [8]. The RTFI selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis. For the specific experiments, an 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, while a sampling rate of 44100Hz is considered for the input samples. 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 discrete RTFI representation 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 version of the real-time adaptive whitening method proposed in [7] is applied, modified for the log-frequency domain. 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.

In addition, a noise suppression approach similar to the one in [4] was employed, due to its computational efficiency. A half-octave span (60 bins) moving median filter is computed for $X[k]$, resulting in noise estimate $N[k]$. Afterwards, an additional moving median filter $N'[k]$ of the same span is applied, but only including the RTFI bins whose amplitude is less than the respective amplitude of $N[k]$. This results in making the noise estimate $N'[k]$ robust in the presence of spectral peaks that could affect the noise estimate $N[k]$.

3. MULTIPLE-F0 ESTIMATION

3.1 Salience Function

A salience function $s[p, d_p, \beta_p]$ is proposed, which indicates the strength of pitch candidates:

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

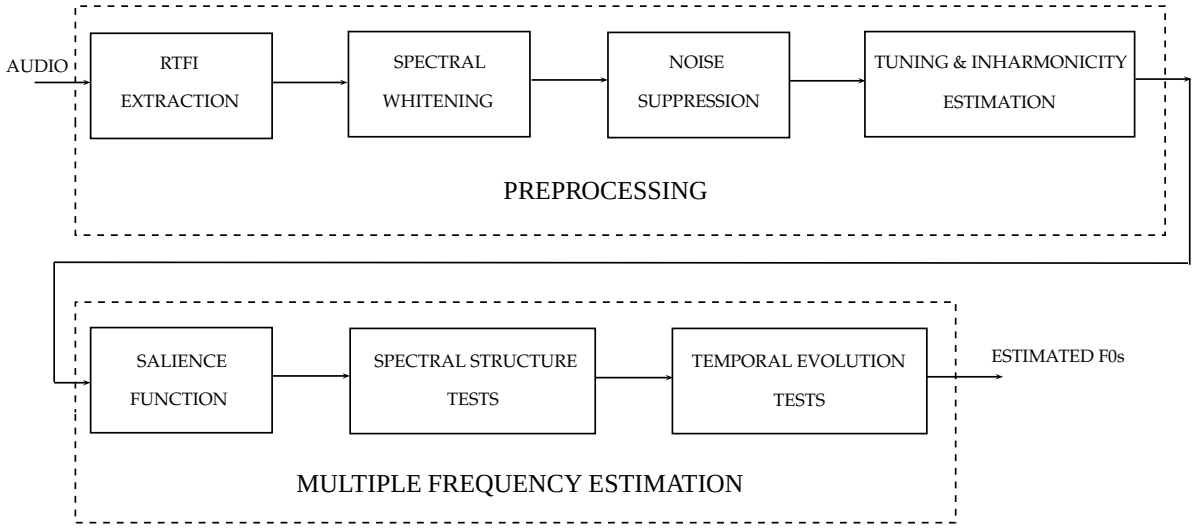

Figure 1. Diagram for the proposed multiple fundamental frequency estimation system.

where

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

where m_h specifies a search range around overtones: $m_h \in \{m_h^l, m_h^u\}$, where $m_h^l = \lceil \frac{\log_2(h-1)+(M-1)\log_2(h)}{M} \rfloor$, $m_h^u = \lceil \frac{(M-1)\log_2(h)+\log_2(h+1)}{M} \rfloor$. $h \geq 1$ is the partial index, β_p is the inharmonicity coefficient [6] and M defines the search range factor, which for the current experiments was set to 60. The number of overtones considered is set to 10 at maximum.

For tuning and inharmonicity estimation, the average spectral representation $\bar{X}[k]$ of the recording is employed. The tuning deviation d_p is set between -40 and +40 cents, centered around the ideal tuning position. The range of the inharmonicity coefficient β_p is set between 0 and $5 \cdot 10^{-4}$, which is typical for piano notes [6]. A two-dimensional maximization procedure is applied to the salience function of $\bar{X}[k]$ for each pitch $p \in \{21, 104\}$ in the MIDI scale. This results to a tuning deviation vector and an inharmonicity coefficient vector. Afterwards, using these settings for d_p and β_p, the salience function (1) is being computed for each frame of $X[k]$, which is now denoted $s'[p]$. From the aforementioned extracted features, a harmonic partial sequence $V[p, h]$ is extracted for each frame in order to be used for further processing (we drop the subscript k for simplicity).

3.2 Spectral Structure Rules

A set of rules aiming to suppress peaks in the salience function that occur at multiples and sub-multiples of the actual fundamental frequencies are applied to $V_k[p, h]$. A first rule for suppressing salience function peaks is setting a minimum number for partial detection in $V[p, h]$, similar to [1, 8]. If $p < 47$, at least three partials out of the first six need to be present in the harmonic partial sequence. If $p \geq 47$, at least four partials out of the first six should be detected. Another processing step in order to reduce processing time is the reduction of the number of candidates, by selecting only the pitches with the greater salience values. In the current experiments, 9 candidate pitches are selected from $s'[p]$.

The spectral flatness [2] is also used for the elimination of errors occurring in subharmonic positions. In the proposed system, the flatness of the first 6 partials of a harmonic sequence is used:

$$Fl[p] = \frac{\sqrt[6]{\prod_{h=1}^{6} V[p, h]}}{\frac{\sum_{h=1}^{6} V[p,h]}{6}} \quad (3)$$

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.

A modified spectral irregularity measure [8] is applied to pairs of harmonically-related candidate F0s (where $f_1 = lf_0$), in order to suppress candidate pitches occurring at multiples of the true fundamental frequency. Given the current set of candidate pitches from $s'[p]$, the overlapping partials from non-harmonically related F0s are detected and smoothed according to the *spectral smoothness* assumption, which states that the spectral envelope of harmonic sounds should form a smooth contour [3]. 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 [3]. The modified spectral irregularity measure is:

$$SI[p, l] = \sum_{h=1}^{3} \frac{2 \cdot V'[p, hl]}{V'[p, h(l-1)] + V'[p, h(l+1)]} \quad (4)$$

For each pair of harmonically-related F0s that are present in $s'[p]$, the existence of the higher pitch is determined by the value of SI' (for the current experiments, a value of 0.6 was set).

3.3 Temporal Evolution Rules

Additional information is exploited in order to produce more accurate estimates in the case of harmonically-related F0s. The *common amplitude modulation* (CAM) assumption [5] is used 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, thus the presence of an additional source that overlaps certain partials causes the correlation between non-overlapped partials and the overlapped partials to decrease.

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:

$$Corr[p,h,l] = \frac{Cov(X[n,k_{p,1}],X[n,k_{p,hl}])}{\sqrt{Cov(X[n,k_{p,1}])Cov(X[n,k_{p,hl}])}} \quad (5)$$

where $k_{p,h}$ indicates the frequency bin corresponding to the h-th harmonic of pitch p, n denotes the RTFI frame number and l the harmonic relation (eg. for octaves $l=2$). Tests are being taken for each pitch p and harmonics hl, using a 5-frame range from the frame under consideration (corresponding to 200ms). 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.

4. RESULTS

For the MIREX task, the system was evaluated using 40 test files from 3 different sources, consisting of several instrument types with maximum polyphony level 5. Results are displayed in Table 1, where it can be seen that the chroma accuracy is increased compared to the system accuracy by 8% (implying octave errors). The system detects very few false alarms and most of the errors consist of missed detections. Overall, the system ranked 4th out of the 8 groups that submitted for the task considering the accuracy measure and 3rd using the chroma acuracy. It should be noted that the system was trained only on piano chords and that no note tracking procedure took place.

	Accuracy	**Precision**	**Recall**
Results	0.468	0.716	0.485
Chroma Results	0.545	0.830	0.567

Table 1. Results for the submitted system.

5. 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] 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, to appear*.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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