# MULTIPLE FUNDAMENTAL FREQUENCY EXTRACTION FOR MIREX 2012

**Karin Dressler**
Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany
kadressler@gmail.com

## ABSTRACT

This extended abstract outlines an efficient approach for the extraction of multiple fundamental frequencies (F0) from polyphonic musical audio. The algorithm consists of three analysis steps. At first a multi-resolution spectral analysis is performed on the audio signal. Then, the most salient pitches are identified using a pitch extraction algorithm, which is designed to identify the predominant pitch in polyphonic audio. Finally, high level tone objects are created and tracked over time: the most salient pitch of the current analysis frame may start a new tone object. All active tone objects are jointly evaluated in order to estimate their pitch and magnitude, and to establish timbre information.

Nonetheless, the MIREX evaluation shows that the system provides excellent results in multiple F0 estimation and tone tracking.

## 1. INTRODUCTION

While note transcription is an important MIR-task in itself, it is also a subtask in many other applications. For example, note transcription can help to improve tempo estimation, melody extraction, or the harmonic analysis of a musical piece. The presented system has been implemented as part of a melody extraction algorithm and therefor places a high priority on the most salient tones and at the processing of a human singing voice. The parameters were tuned in respect to the best melody extraction results, so the used setting is probably not the best choice to maximize the estimation accuracy for the multiple F0 task – in particular, as the dataset for melody extraction consists mostly of musical pieces with a singing voice, while the multiple F0 dataset includes solely instrumental music.

Three distinct algorithms have been submitted to MIREX: two algorithms providing the frequencies of all extracted tones sampled at a 10 ms interval (multiple F0 estimation) and one algorithm which outputs the onset and offset time as well as the MIDI note number for each extracted note (tone tracking).



## 2. METHOD

The presented algorithm was implemented as part of a melody extraction algorithm, which was evaluated at the Music Information Retrieval Evaluation eXchange (MIREX) in 2009 [1]. However, there is one modification for the multiple F0 estimation task: the frequency range for tones was increased to cover frequencies between 55 Hz and 2093 Hz.

### 2.1 Spectral Analysis and Magnitude Weighting

If a partial of a complex tone is not obscured by other harmonics or noise, it can be detected as a peak in the magnitude spectrum of the Short Term Fourier Transform (STFT). The interference of partials from simultaneously playing notes can be decreased if the frequency resolution of the STFT is increased. However, musical sound is not stationary, so very long STFT data windows cannot be used to gain a very high frequency resolution. As a compromise between a good frequency resolution and a good time resolution, we analyze the audio signal by calculating a multi-resolution Fast Fourier Transform (MR FFT) [2].

The best frequency resolution ($\Delta f = 21.5$ Hz) is reached for the low frequency components up to approximately 600 Hz. The best time resolution corresponds to a FFT data window length of 5.8 ms for frequencies above 4400 Hz. Due to different amounts of zero padding the resulting STFT frame length and the hop size of the analysis window correspond to 46 ms and 5.8 ms for all STFT resolutions.

In order to obtain the weighted magnitude $A_s$ for the spectral peak at STFT bin $k$, its STFT magnitude is multiplied with the peak's instantaneous frequency $f_i$.

$$A_s[k] = |X[k]| \cdot f_i[k] \tag{1}$$

This weighting introduces a 6 dB magnitude boost per octave. In effect, the weighted signal is proportional to the signal derivative.

### 2.2 Pitch Estimation

For the computation of the pitch spectrogram, spectral peaks in the frequency range between 55 Hz and 5 kHz are processed. The weighted magnitude and the instantaneous frequency of the spectral peaks are evaluated in order to identify the strongest signal periodicity in the frequency range between 55 Hz and 2093 Hz. The pitch estimation algorithm is based on the pair-wise analysis of spectral peaks [3]. The idea of the technique lies in the identification of partials with successive (odd) harmonic numbers. Since successive partials of a harmonic sound have well defined frequency ratios, a possible fundamental frequency (F0) can be derived from the instantaneous frequencies of two spectral peaks. Consecutively, the identified harmonic pairs are rated according to harmonicity, timbral smoothness, the appearance of intermediate spectral peaks and harmonic number. Finally, the resulting pitch strengths are added to a pitch spectrogram.

### 2.3 Tones

A high level tone object is started, if the most salient pitch in the current analysis frame passes an adaptive magnitude threshold.

All active tone objects are jointly evaluated over time in order to estimate their pitch and their magnitude. At the same time a spectral envelope is established for each tone. The spectral envelope (e.g. harmonic magnitudes) determines the weight each spectral peak receives in the tone's pitch and magnitude estimation. In this way, the impact of noise and concurrent tones can be decreased noticeably.

In order to establish long term timbre information, adequate spectral peaks are assigned to the active tone objects in each analysis frame. The added spectral peaks, eventual masking and the computed tone height are exploited in a rating scheme that determines how well each harmonic can be integrated into the overall timbre. The principle indicators for the harmonic fit are: 1) the frequency difference between tone height and computed virtual pitch of the harmonic, 2) the smoothness of the timbre in the frequency and time dimension, and 3) the magnitude division of shared harmonics among distinct tones.

A feedback about the existing tone objects is provided to the pitch determination method, so that matched spectral peaks can be inhibited during the pitch determination. This way, pitches besides the predominant pitch can be extracted.

## 3. EVALUATION

Three algorithms have been submitted: KD1 and KD2 are multiple F0 estimation algorithms that detect the occurring F0 in each analysis frame. Compared to KD1 there are some modifications in KD2 to improve the onset and offset estimation of tones. The MIREX results show that these modification have no significant impact on the frame-wise estimation. Algorithm KD3 is based on submission KD2, yet has a different output format, as it does give note onset, note offset and the perceived tone height [1].

### 3.1 Task 1: Multiple Fundamental Frequency Estimation

40 test files were analyzed for this task: 20 excerpts from the woodwind recording recording of bassoon, clarinet, horn, flute and oboe (polyphony ranging from 2 to 5), 12 excerpts from a quartet recording of bassoon, clarinet, violin and sax (polyphony ranging from 2 to 4), and 8 files from synthesized MIDI (polyphony ranging from 2-5).

#### *3.1.1 Evaluation Metrics*

A pitch estimate is assumed to be correct if it is within a half semitone ($\pm 50$ cent) of a ground-truth pitch for that frame. Only one ground-truth pitch can be associated with each returned Pitch. Two different sets of evaluation metrics are used to estimate the algorithm performance. The first set estimates the performance in terms of precision, recall and overall accuracy using the following equations:

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

$$Accuracy = \frac{TP}{TP + FP + FN}, \quad (4)$$

where $TP$ is the number of correctly identified pitches (true positives), $FP$ is the number of identified pitches which do not occur in the ground truth (false positives), and $FN$ is the number of pitches which are not identified by the algorithm (false negatives).

The second set of evaluation metrics was proposed by Poliner and Ellis in order to measure the accuracy of polyphonic piano transcriptions [4]. The metric computes an error score $E_{\text{tot}}$ that takes into account the so-called substitution errors $E_{\text{subs}}$, which allow the substitution of any false positive F0 with a missing ground-truth F0 [4]. The number of errors is set into relation to the total quantity of notes:

$$E_{\text{subs}} = \frac{\sum_{t=1}^{T} \min(N_{\text{ref}}(t), N_{\text{sys}}(t)) - N_{\text{corr}}(t)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}, \quad (5)$$

where $N_{\text{ref}}$ is the number of pitches in the ground truth data, $N_{\text{sys}}$ is the number of pitches returned by the system, $N_{\text{corr}}$ is the number of correctly identified pitches, and $t$ is the index of the current analysis frame.

The other components of the metric are missing pitches $E_{\text{miss}}$ and false alarm errors $E_{\text{fa}}$. While $E_{\text{miss}}$ refers to the number of ground-truth reference notes that could not be matched with any system output (i.e. misses after substitutions are accounted for), $E_{\text{fa}}$ refers to the number of pitches that cannot be paired with any ground truth (false alarms beyond substitutions):

$$E_{\text{miss}} = \frac{\sum_{t=1}^{T} \max(0, N_{\text{ref}}(t) - N_{\text{sys}}(t))}{\sum_{t=1}^{T} N_{\text{ref}}(t)} \quad (6)$$

$$E_{\text{fa}} = \frac{\sum_{t=1}^{T} \max(0, N_{\text{sys}}(t) - N_{\text{ref}}(t))}{\sum_{t=1}^{T} N_{\text{ref}}(t)}. \quad (7)$$

The total error is estimated as follows:

$$E_{\text{tot}} = \frac{\sum_{t=1}^{T} \max(N_{\text{ref}}(t), N_{\text{sys}}(t)) - N_{\text{corr}}(t)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}. \quad (8)$$

[1] More detailed information about the MIREX multiple fundamental frequency estimation task and the results can be found online at: `http://www.music-ir.org/mirex`

| | Precision | Recall | Accuracy | Etot | Esubs | Emiss | Efa | Runtime (sec) |
|---|---|---|---|---|---|---|---|---|
| BD1 | 0.64 | 0.72 | 0.58 | 0.56 | 0.16 | 0.12 | 0.28 | tba |
| CPG1 | 0.58 | 0.28 | 0.27 | 0.74 | 0.19 | 0.53 | 0.02 | tba |
| CPG2 | 0.58 | 0.27 | 0.27 | 0.75 | 0.2 | 0.53 | 0.02 | tba |
| CPG3 | 0.57 | 0.27 | 0.27 | 0.75 | 0.2 | 0.53 | 0.02 | tba |
| FBR1 | 0.58 | 0.88 | 0.56 | 0.76 | 0.08 | 0.04 | 0.65 | tba |
| KD1 | 0.84 | 0.66 | 0.64 | 0.38 | 0.09 | 0.25 | 0.04 | tba |
| KD2 | 0.86 | 0.67 | 0.64 | **0.38** | 0.08 | 0.26 | 0.04 | tba |
| YR2 (2011) | 0.73 | 0.84 | **0.68** | 0.42 | 0.08 | 0.08 | 0.26 | 6584 |

**Table 1**. Task 1: Multiple Fundamental Frequency Estimation Results

| | BD2 | BD3 | CPG1 | CPG2 | CPG3 | FBR2 | FT1 | KD3 | SB5 |
|---|---|---|---|---|---|---|---|---|---|
| Ave. F-Measure Onset-Offset | 0.23 | 0.23 | 0.13 | 0.13 | 0.11 | 0.39 | 0.02 | **0.45** | 0.09 |
| Ave. F-Measure Onset Only | 0.43 | 0.41 | 0.22 | 0.23 | 0.27 | 0.61 | 0.06 | **0.65** | 0.5 |
| (piano only) | | | | | | | | | |
| Ave. F-Measure Onset-Offset | 0.13 | 0.19 | 0.14 | 0.14 | 0.16 | 0.16 | 0.07 | **0.27** | 0.04 |
| Ave. F-Measure Onset Only | 0.5 | 0.61 | 0.3 | 0.31 | 0.38 | 0.62 | 0.17 | **0.66** | **0.66** |

**Table 2**. Task 2: Tone Tracking Results

#### *3.1.2 Results and Discussion*

Table 1 shows the results for the frame-wise multiple F0 estimation. The results of our two submissions KD1 and KD2 do not differ significantly, but compared with the result of last year, the accuracy of our algorithm has improved by one percent. The accuracy (64%) marks the best result in this year, but it does not reach the accuracy of the system by Yeh and Roebel, which was submitted to MIREX 2011 [5]. Then again, our system performs best in terms of the total error metric $E_{\text{tot}}$ introduced by Polliner and Ellis in [4].

It can also be noted that on the one hand, our algorithm has the highest precision (i.e. 86% of the extracted fundamental frequencies are true positives), but on the other hand, it systematically underestimates the number of concurrent voices, leading to a low recall (67%). However, this fact is not very surprising, as the main purpose of a melody extraction algorithm is to detect the strongest notes of a musical piece, and not a transcription of all notes. While a better trade-off between precision and recall might be achieved by using a more suitable dataset for the parameter estimation, it may not be possible to reach a much better result without loosing some generality in terms of the input data.

### 3.2 Task 2: Note Tracking

A total of 34 files were analyzed in this subtask: 16 excerpts from woodwind recordings, 8 excerpts from the IAL quintet recording and 6 piano recordings.

#### *3.2.1 Evaluation Metrics*

For this task the F-Measure is reported, which is the harmonic mean of precision and recall (see equations 2 and 3) for each input file:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{9}$$

Then the average is calculated from the results of the individual files.

A ground truth note is assumed to be correctly transcribed if the transcription system returns a note that is within a half semitone of that note AND the returned note onset is within a 100 ms range($\pm 50$ ms) of the onset of the ground truth note, and its offset is within a $20\%$ range of the ground truth note offset. The evaluation of the note offset is omitted in the "onset-only" subtask.

#### *3.2.2 Results and Discussion*

Reaching an average F-measure of 0.45, the algorithm KD3 marks the state of the art in note tracking (see table 2).

In general, it is much easier to detect note onsets than note offsets – a fact that is particularly apparent in the piano dataset, where all algorithms suffer from bad offset detection results.

If we take a look at the onset-only piano subtask, it is very encouraging to see that the accuracy of our system does not differ significantly from the result achieved by submission SB5. This is remarkable, as the latter system (which is based on a recurrent neural network that was explicitly trained for piano note onset transcription) marks the state of the art in this specific task.

## 4. CONCLUSION

In this extended abstract we presented an efficient approach to the estimation of multiple F0 from polyphonic music. The MIREX results show that the proposed method allows not only a reliable and very efficient identification of the fundamental frequencies in each analysis frame, but also succeeds in the formation of continuous tone objects.

## 5. REFERENCES


[1] K. Dressler. Audio Melody Extraction for MIREX 2009. In *5th Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

[2] K. Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, pages 247–252, Montreal, Quebec, Canada, Sept. 2006.

[3] K. Dressler. Pitch estimation by the pair-wise evaluation of spectral peaks. In *AES 42nd Conference*, Ilmenau, Germany, July 2011.

[4] G.E. Poliner and D.P.W. Ellis. A discriminative model for polyphonic piano transcription. In *EURASIP Journal on Advances in Signal Processing*, 2007

[5] C. Yeh and A. Roebel Multiple-F0 estimation for MIREX 2011. In *7th Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.