REDUCING OCTAVE ERRORS IN TEMPO ESTIMATION BY EMPLOYING HARMONIC CONTENT CHANGES

Taehoon Kim, Kyogu Lee

*Music and Audio Research Group, Graduate School of Convergence Science and Technology, Seoul National University, Seoul 151-742, Korea*

e-mail: kcjs55@snu.ac.kr

Conventional tempo estimation algorithms attempt to detect a global tempo by means of searching noticeable acoustic events or periodicities of repetitive patterns in a music signal. However, musical rhythm has a hierarchical structure in nature and such energy-based algorithms for tempo estimation suffer from a problem of mis-estimating the tempo twice as fast or as slow as the original tempo, known as a double-time/half-time or octave problem. In this paper, we propose a precise tempo estimation algorithm to alleviate the octave errors by taking into account other musical information in an audio signal, especially the changes in the harmonic content. To this end, we first estimate a set of initial tempi using the conventional tempo detection method based on the periodicity of the novel acoustic events. Then, we find the boundaries where the harmonic content significantly changes over time, and calculate the histogram of the durations between the adjacent boundaries. This allows us to statistically find the representative rate of the harmonic changes, and to use it to select the most appropriate tempo among the candidate tempi. Experimental results show that the proposed method effectively reduces the octave errors in tempo estimation for the various types of musical signals. Moreover, we also show that the proposed tempo estimation algorithm helps find the precise locations of the musical beats as human perceives, which is a difficult task for many conventional beat-tracking algorithms.

1. Introduction

A tempo, defined as speed of a given piece in musical terminology, is an important element in musical compositions or auditory cognition since it is closely concerned with mood or atmosphere of the piece. To detect a tempo of an audio signal automatically, many tempo estimation algorithms have been so far proposed. Most conventional tempo estimation algorithms detect a global tempo by searching acoustic events or periodicities of repetitive patterns in a music signal. However, most of them are suffering from an octave error problem, and fail to find a correct tempo (i.e., ground-truth tempo) but detect fractions or multiples of the correct tempo.

In this paper, we propose a novel tempo estimation scheme to reduce the octave error problem based on the observation that a tempo is closely related with the changes in harmonic contents as well as periodicities in acoustic energy of musical audio. To this end, we first estimate a provisional tempo with conventional tempo estimation algorithms and create a candidate tempo set with the detected tempo. To decide a final tempo among the tempo candidates, the proposed method extracts harmonic contents from the audio signal. Then, we define a new hybrid function that includes information on harmonic content changes and an envelope of an onset detection function. Dominant peaks in an
autocorrelation function of this newly created function indicate strong candidates for the final tempo. Lastly, we determine the final tempo by comparing the most dominant peak of the autocorrelation function with candidate tempi in the candidate set. Experimental results show the effectiveness of the proposed algorithm by reducing the octave errors and thus enhancing the accordance rate between the detected tempo and ground-truth tempo.

The remainder of this paper is organized as follows. Section 2 describes the related works on tempo estimation and octave errors in particular. We explain in detail the proposed tempo estimation method in Section 3. We present the experimental results in Section 4 and draw conclusions in Section 5.

2. Related work

2.1 Tonal centroid

Understanding tonality is an important factor to analyze western tonal music, explaining the relationship among the tones. Constant-Q transform transforms an audio signal to the frequency domain with logarithmically spaced filters, and the center frequency of the \( k \)-th filter \( f_k \) is given by

\[
f_k = (2^{\frac{Q}{B}}) \cdot f_{k-1} = (2^{\frac{Q}{B}})^k \cdot f_{\min}, \quad k = 0, \cdots, N - 1,
\]

where \( B \) is the number of bins in an octave and \( f_{\min} \) is the minimum frequency. Then, the audio signal can be transformed using constant-Q transform as follows:

\[
X_{eq}(n) = \frac{1}{N(k)} \sum_{n=0}^{N(k)} x(n) w(n, k) e^{-j2\pi Qn/N(k)}.
\]

\( N(k) \) is defined as \( \frac{f_Q}{f_k} \), where \( Q \) is the quality factor and \( f_s \) is the sampling rate of the audio signal. Then chroma, also known as pitch class profile, can be computed from the constant-Q transform as follows:

\[
Chroma(b) = \sum_{m=0}^{M-1} \left| X_{eq}(b + mB) \right|, \quad b = 0, \cdots, B - 1.
\]

Harte and Gasser proposed a tonal centroid for better chord recognition by using a concept of harmonic network. Tonal centroid is 6-dimensional feature linearly transformed from 12-dimensional chroma feature (i.e., \( Chroma \) with \( B = 12 \)) and equation 4 describes derivation procedure of tonal centroid.

\[
TC(d) = \frac{1}{\Phi(Chroma(b))} \sum_{b=0}^{B-1} \Phi(d, b) Chroma(b), \quad d = 0, \cdots, 5
\]

where \( \Phi = [\phi_0, \cdots, \phi_{B-1}] \) is 6x12 transformation matrix, where \( \phi_b \) is defined as

\[
\phi_b = \begin{bmatrix}
r_1 \sin(b \frac{7\pi}{6}) & r_1 \cos(b \frac{7\pi}{6}) & r_2 \sin(b \frac{3\pi}{2}) & r_2 \cos(b \frac{3\pi}{2}) & r_3 \sin(b \frac{2\pi}{3}) & r_3 \cos(b \frac{2\pi}{3})
\end{bmatrix}^T.
\]
2.2 Tempo estimation algorithm using autocorrelation function

The tempo estimation algorithm employing autocorrelation function of detection functions is proposed by Tzanetakis and Cook. This algorithm is considered in this paper as a conventional scheme to figure out initial tempo. It operates on onset detection function and uses subband analysis.

To be specific, it first decomposes the audio signal into multiple octave frequency bands using discrete wavelet transform. Following this process, the time domain amplitude envelope of each band is extracted respectively. This can be done by employing low-pass filtering, full-wave rectification and down-sampling to all of frequency bands. Then, envelopes of every band are added and the autocorrelation function of the result is calculated using following equation,

$$y(k) = \frac{1}{N} \sum_{n} f_{env}(n) f_{env}(n-k).$$

Dominant peaks of the autocorrelation function correspond to the periodicities of the summed envelope and are accumulated as beat intervals, which can be converted into tempi in beat-per-minutes (BPM) metric.

3. Proposed method

We consider the reduction of octave errors in tempo estimation schemes by considering changes of harmonic contents in the audio signal. As illustrated in Fig. 1, each child device can only utilize one of the partitioned sub-transmission periods for the packet transmission. We consider three main steps; candidate tempo set generation, harmonic change interval estimation and final tempo decision. Figure 1 shows overall procedure of proposed tempo estimation method.

![Figure 1. Overall procedure of proposed tempo estimation method](image-url)
3.1 Generating a candidate tempo set

To identify a correct tempo, first we need to detect an initial tempo and correct it with a harmonic change detect function which will be described in subsection 3.2. When detect an initial tempo, any conventional tempo estimation scheme will do. However, since accuracy of the initial tempo affects correctness of a final tempo (though the initial tempo has octave error problem), selection of tempo estimation scheme is important.

By the use of initial tempo, proposed scheme generates a candidate tempo set. One of components in the candidate tempo set will be decided as a final tempo. The candidate tempo set contains multiples and divisors of the initial tempo. For instance, a candidate tempo set $\Theta$ can be determined as,

$$\begin{align*}
\Theta &= \left\{ \frac{1}{6}T, \quad \frac{1}{4}T, \quad \frac{1}{3}T, \quad \frac{1}{2}T, \quad \frac{2}{3}T, \quad \frac{3}{2}T, \quad 2T \right\}.
\end{align*}$$

(7)

A final candidate tempo set can be customized according to the tempo estimation scheme used to calculate the an initial tempo and we suggest that components are form of $2k \cdot 3^l \cdot T$, $k \leq 1$, $l \leq 1$.

(8)

In this manner, chosen final tempo can solve double/triple-tempo problems.

3.2 Harmonic change interval estimation

Assume that $x[n]$ is a primitive audio signal in temporal dimension and firstly the audio signal is converted into the spectrogram $S(k, n_j)$ as,

$$S(k, n_j) = \sum_{n=0}^{W-1} x(n + n_j) \cdot e^{-j \frac{2\pi n}{N}}, \quad k = 0, \ldots, W - 1$$

(9)

where $k$, $n_j$ represents spectral/temporal index of spectrogram respectively and $W$ is window size of fast Fourier transform(FFT). Since $W$ influences resolution of spectrogram, which have a direct effect on calculating tonal centroids of spectrogram, $W$ cannot be randomly chosen. As $W$ increases, probability of detecting long intervals between adjacent harmonic changes becomes larger. For this reason, proposed scheme determines $W$ as,

$$W = \mu \cdot f_s \cdot T$$

(10)

where $f_s$ is a sampling rate of the primitive audio signal and $\mu$ is a compensation factor. Tonal centroid is calculated as follows,

$$\zeta_{n_j}(d) = \frac{1}{\|c_{n_j}\|} \sum_{j=0}^{11} \Phi(d, b)c_{n_j}(b), \quad 0 \leq d \leq 5, \quad 0 \leq b \leq 11$$

(11)

where $c_{n_j}$ is a 12-dimension chroma vector on time-frame $n_j$ calculated with $S(k, n_j)$ and $\Phi$ is a $6 \times 12$ transformation matrix defined in section 2. Then $\zeta_{n_j}$ shows harmonic contents in the audio signal on $n_j$ and harmonic content changes can be derived by calculating differentials of tonal centroid.
values in previous and next time-frame. In this manner, a harmonic content change detection function (HCDF) $\delta_{n}$ can be calculated as,

$$\delta_{n} = \left( \sum_{d=0}^{5} (\zeta_{n+d} - \zeta_{n-d}) \right)^{2}.$$  

(12)

It is necessary to find peaks of HCDF to detect changes of harmonic content, therefore a peak-picking process is required. Because of roughness of HCDF, we adopt an adaptive threshold to locate peaks. The adaptive threshold can be calculated as follows,

$$\varphi_{n} = \alpha + \beta \cdot \text{med}\left(\delta(k)\right), \quad k \in \left[ t - \frac{H}{2}, t + \frac{H}{2} \right]$$  

(13)

where $\alpha$, $\beta$ is constant, med( ) means a median value of the function displayed in parenthesis and $H$ is a window size around the time frame deciding the number of samples to determine the threshold. Samples which satisfy a condition $\delta_{n} > \varphi_{n}$ are picked as peaks and proposed scheme defines a new function $f_{h}$ which is sum of the envelope and HCDF, calculated as,

$$f_{h}(n) = f_{env}(n) + \sigma \cdot \frac{\text{max}(f_{env})}{\text{max}(\delta)} \delta_{n}.$$  

(14)

where max( ) means a maximum value of a function in the parenthesis and $\sigma$ represents a correction factor. The correction factor can be chosen adaptively according to features of music, however, in this paper, the value is fixed with a specific constant value. Function $f_{p}$ represents a hybrid function containing information of onsets and harmonic content changes simultaneously. Peaks picked from the hybrid function show the beat interval of the audio signal considering both factors as humans do.

To find intervals between adjacent peaks, an autocorrelation function of $f_{p}$ is calculated as explained in section 2.2,

$$C(k) = \frac{1}{N} \sum_{n} f_{h}(n) f_{h}(n-k),$$  

(15)

We decide the time value of the largest local maximum point except the point $\tau = 0$ as a harmonic content change interval. An equation 10 shows a process of concluding the value of the harmonic content change interval.

$$\text{maximize}_{k} \quad C(k)$$  

subject to $$\frac{dC(k)}{dk} = 0, \quad k > 0.$$  

(16)

### 3.3 Final tempo decision

To compare the harmonic content change interval with tempi in the candidate tempo set derived in sub-section 4.1, an unit of the harmonic content change interval should be converted into beat per minute(BPM),

$$\nu_{HCl} = \frac{60}{T_{HCl}}.$$  

(11)
As mentioned in sub-section 4.1, the harmonic content change interval and components in candidate tempo set are compared and decide one which is closest to the harmonic content change interval as a final tempo of music. By employing equation 12, we can decide an index of final tempo among components.

\[
k_{\text{final}} = \arg\min_k (\theta_k - \nu)_{\text{HCl}}, \quad k = 1, \ldots, |\Theta|
\]

(12)

where \(|\Theta|\) describes a cardinality of set \(\Theta\) and \(\theta_k\) represents a \(k\)-th element of set \(\Theta\). \(k_{\text{final}}\) is an index of the element with closest value to the harmonic content change interval. Then, a final tempo can be decided as,

\[
T_{\text{final}} = \theta_{k_{\text{final}}}
\]

(13)

4. Performance evaluation

For evaluation, two sets of music data were considered: the MIREX practice dataset\(^4\) and the Beatles annotations\(^5\). Since they contain ground-truth tempo information with audio files and are open to the public, we chose to use these datasets to evaluate performance of the proposed method. For comparison, the performance of conventional tempo estimation algorithm was also evaluated.

![Figure 2. Detected tempi of MIREX practice dataset](image)

Figure 2 illustrates the performance comparison for the MIREX practice dataset. MIREX practice set is composed of 20 songs with various genres, each clip being 30 seconds long. It can be seen
in the figure that the tempi estimated by the conventional algorithm do not correspond well to the ground-truth tempi. In many cases, detected tempi using the conventional method show double or triple of ground-truth tempi, or octave errors. However, by using the proposed algorithm, octave errors are considerably reduced as depicted, mainly due to the fact that it also considers harmonic contents of the audio signal and determines tempi as humans do. But still, as shown in cases of a 20-th song in the graph, proposed scheme rather shows a wrong result since the proposed scheme considers the correction factor $\sigma$ identically for every piece regardless of characteristics of pieces. Moreover, for a few cases, a confusion between double-tempo and triple-tempo is occurred (in case of song 15).

Table 1 displays accuracies in % of the two algorithms for the Beatles dataset. Correctness of detection is defined as a coincidence between detected tempi and ground-truth tempi in each album. The results show a significant improvement in accuracy with the proposed algorithm (approximately 82.2% increase in performance from 44.6% to 81.3%).

Table 1. Performance comparison for each album in the Beatles dataset (%)

<table>
<thead>
<tr>
<th>Album no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv.</td>
<td>64.28</td>
<td>28.57</td>
<td>50.00</td>
<td>57.14</td>
<td>42.85</td>
<td>78.57</td>
<td>46.15</td>
<td>27.27</td>
<td>17.65</td>
<td>33.33</td>
<td>44.58</td>
</tr>
<tr>
<td>Prop.</td>
<td>85.71</td>
<td>92.86</td>
<td>100.00</td>
<td>78.57</td>
<td>92.86</td>
<td>92.86</td>
<td>84.62</td>
<td>72.73</td>
<td>70.99</td>
<td>41.67</td>
<td>81.25</td>
</tr>
</tbody>
</table>

In the Beatles annotations, no information about a dominant tempo is given but beat positions. Therefore, we calculated inter-beat intervals using the beat position information and computed ground-truth tempo. Albums in consideration are listed below:

- Album 1: “Please Please Me”, 1963, 14 tracks
- Album 2: “With the Beatles”, 1963, 14 tracks
- Album 3: “Beatles for sale”, 1964, 14 tracks
- Album 4: “Help!”, 1965, 14 tracks
- Album 5: “Rubber Soul”, 1965, 14 tracks
- Album 6: “Revolver”, 1966, 14 tracks
- Album 8: “Magical Mystery Tour”, 1967, 11 tracks
- Album 9: “Abbey Road”, 1969, 17 tracks

5. Conclusion

In this paper, we have proposed a method to reduce octave errors in tempo estimation by detecting harmonic contents of an audio signal. The hypothesis was that the tempo of given musical audio is determined not only by the periodic events but also by the changes in harmonic contents in
a signal. To this end, we first estimate a set of initial tempi using the conventional tempo detection method based on the periodicity of the novel acoustic events. Then, we find the boundaries where the harmonic content significantly changes over time, and calculate the histogram of the durations between the adjacent boundaries. This allows us to statistically find the representative rate of the harmonic changes, and to use it to select the most appropriate tempo among the candidate tempi.

In the proposed method, the correction factor or $\sigma$ has proven to have a critical effect on the final results. We plan in the near future to improve the performance by determining an appropriate correction factor $\sigma$ using machine-learning approaches considering acoustic features.

ACKNOWLEDGEMENT

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-H0301-13-4005).

REFERENCES