A REAL-TIME FRAME-BASED MULTIPLE PITCH ESTIMAITON METHOD USING THE RESONATOR TIME-FREQUENCY IMAGE

Ruohua Zhou
Center for Digital Music,
Electronic Engineering Department,
Queen Mary University
ruohua.zhou@elec.qmul.ac.uk

Joshua D Reiss
Center for Digital Music,
Electronic Engineering Department,
Queen Mary University
josh.reiss@elec.qmul.ac.uk

ABSTRACT

In this paper, we describe a new method submitted to the 2007 Music Retrieval Evaluation eXchange (MIREX) Multiple Fundamental Frequency Estimation task. The Resonator Time-Frequency Image (RTFI) is used as the basic time-frequency analysis tool. The method makes a preliminary pitch estimation by a simple peak-picking operation in the pitch energy spectrum, which is produced from the original energy spectrum according to harmonic grouping principle. Then the incorrect estimations are cancelled according to some assumptions about harmonic structures of music notes played by commonly-used music instruments. The evaluation of this method under the MIREX task is also discussed.

1. INTRODUCTION

Multiple pitch estimation has a wide range of application. The purpose of this paper is to describe a method for real-time frame-based multiple pitch estimation. In the method, an estimation and cancellation scheme is employed. First, a preliminary estimation is expected to find all possible pitch candidates, which may include extra incorrect estimations. Then, the extra estimations are cancelled. The method uses the Resonator Time-Frequency Image (RTFI) as the time-frequency representation.

2. TIME-FREQUENCY ANALYSIS

2.1. Resonator Time-Frequency Image

The Resonator Time-Frequency Image (RTFI) is a computationally efficient time-frequency representation for music signal analysis. The RTFI selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis. Using the RTFI, one can select different time-frequency resolutions, such as uniform analysis, constant-Q analysis, or ear-like analysis by simply setting different parameters; and letting the RTFI generalize all these analyses in one framework. In this paper, the multi-resolution fast RTFI has been used for time-frequency analysis. The more detailed description of the RTFI can be found in [1] and [2].

2.2. Multi-resolution Fast RTFI

The basic idea of the Multi-resolution Fast RTFI is to reduce the redundancy in computation. In some cases it is not necessary to keep the same sampling frequency of the input for every filter in the filter bank. For the filters with lower center frequencies, the sampling rate can be decreased.

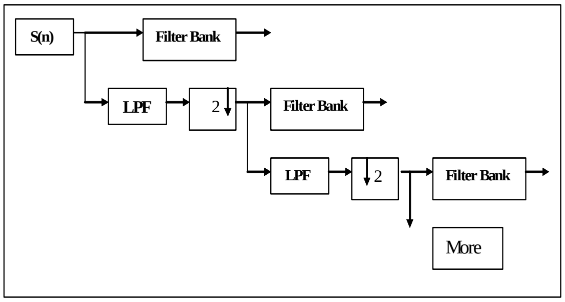

Figure 1. Block diagram for the multi-resolution fast implementation

In the fast implementation, the filter bank is separated into different octave frequency bands. The inputs of the filter banks in same frequency band keep the same sampling rate. The input signal is recursively low-pass filtered and down sampled by a factor 2 from the highest to the lowest frequency band according to the scheme depicted in Figure 1.

2.3. Specific Implementation

In the first step of the method, the RTFI is used to analyze the input music signal and to produce a time-frequency energy spectrum. The input sample is a monaural music signal frame at a sampling rate of 44.1kHz. The frame length is not necessarily fixed. This is designed for a real polyphonic music transcription task, in which the frame length varies for different music notes. Multi-resolution fast RTFI with constant-Q resolution is selected for the time-frequency analysis to produce the time-frequency energy spectrum. All 1050 filters are used. The centre frequencies are set in logarithmic scale. The centre frequency difference between two neighbouring filters is equal to 0.1 semitone and the analyzed frequency range is from 46Hz to 11 kHz. Then, the time-frequency energy spectrum is averaged for every 10ms frame. This RTFI average energy spectrum is used as the only input vector for the multiple pitch estimation method.

Let us define the RTFI average energy spectrum as follows,

$$ARTFI\ (l,\omega_m) = db(\frac{1}{M}\sum_{i=(l-1)M+1}^{lM} \left|RTFI\ (n,\omega_m)\right|^2) \tag{1}$$

where M is an integer and the ratio of M to sampling rate is the duration time of the frame in the average process. In this paper, M is set to 441 corresponding to the frame duration time of 10ms. $RTFI(n, \omega_m)$ denotes the value of discrete RTFI at sampling point n and frequency ω_m, l is the index of frame. The multiple pitches are estimated for every frame.

3. MULTIPLE PITCH ESTIMAITON METHOD

The method is based on estimation and cancelling schema. The multiple pitches are initially estimated, and then the incorrect estimations are cancelled.

3.1. Preliminary Estimation

Based on the harmonic grouping principle, the input RTFI energy spectrum is first transformed into the pitch energy spectrum (PES) and the relative pitch energy spectrum (RPES) as follows:

$$PES\ (f_k) = \frac{1}{L}\sum_{i=1}^{L} ARTFI\ \ (i \cdot f_k) \tag{2}$$

$$RPES(f_k) = PES(f_k) - \frac{1}{N_2+1}\sum_{i=k-N_2/2}^{k+N_2/2} PES(f_i) \tag{3}$$

Where L is a parameter that denotes how many low harmonic components are together considered as important evidence for judging the existence of a possible pitch.

The ideal parameter L and N_2 value need to be set by the experiments on the tuning database. In the following reported test experiments, L and N_2 are fixed at 4 and 50 respectively.

In the description of this method, the integer k is used to denote the frequency index in the logarithmic scale, whereas f_k denotes the corresponding frequency value in Hz as follows:

$$f_k = 440 \cdot 2^{(k-690)/120} \tag{4}$$

In practical implementations, instead of using the equation (2) the pitch energy spectrum can be easily approximated in the logarithmic scale by the following calculation (here L is less than 10):

$$PES\ (f_k) = \frac{1}{L}\sum_{i=1}^{L} RTFI\ (f_{k+A[i]}) \tag{5}$$

$$A[10] = [0,120,190,240,279,310,337,360,380,399]$$

i	1	2	3	4	5
$_{i}\,\frac{f_{k+A[i]}}{i \cdot f_k}$	0%	0%	-0.1%	0%	0.2%

Table 1. Deviation between approximation and ideal values

As shown in Table 1, the deviation between the approximate and ideal values is negligible.

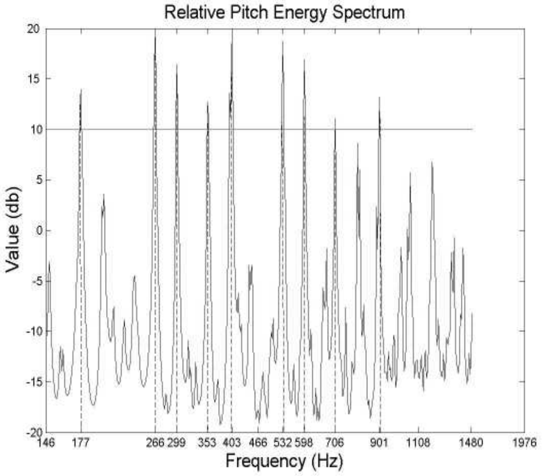

Figure 2. Relative Pitch Energy Spectrum of a violin example consisting of four concurrent notes with the fundamental frequencies 266Hz, 299Hz, 353Hz and 403Hz.

The transformation from the original energy spectrum to a relative energy spectrum has been proven by experiments to be very useful for improving the method's performance. The preliminary estimates of the possible pitches are based on the relative pitch energy spectrum using the following assumption. If there is a pitch with fundamental frequency $f_{k,}$ in the input sample, there should be a peak around the frequency f_k in the relative pitch energy spectrum, and the peak value should surpass a threshold A_2.

Figure 2 illustrates the relative pitch energy spectrum of a violin example, which consists of four concurrent notes with fundamental frequencies of 266Hz, 299Hz, 353Hz and 403Hz respectively. As shown in Figure 4, there are 9 pitch candidates that can be preliminarily estimated. when selecting the threshold A_2= 10db, The fundamental frequencies of the 9 pitch candidates are 177 Hz, 266Hz, 299Hz, 353Hz, 403Hz, 532Hz, 598Hz, 796Hz and 901Hz. In this preliminary estimate, all 4 true pitches in the example have been correctly estimated. On the other hand, 5 extra pitches have been incorrectly estimated. The estimated extra pitches usually share many harmonic components with the true pitches. In this example, the estimated extra pitch of 177Hz is nearly half the value of the true pitch of 353 Hz. The extra pitches 532Hz and 598Hz are nearly twice that of the true pitch 266Hz and the true pitch 299Hz respectively, and the extra pitch 796Hz is nearly triple of the true pitch 266Hz.

3.2. Cancelling Extra Estimation

The input RTFI energy spectrum can be transformed to the relative energy spectrum (RES) according to the following expression:

$$RES\,(f_k) = RTFI\,(f_k) - \sum_{i=k-N_1/2}^{k+N_1/2} RTFI\,(f_i) \tag{6}$$

k=1,2,3,…

where the second term in the right hand part of the equation denotes the moving average of the input RTFI energy spectrum, and N_1 is the length of the window for calculating the moving average.

If there is a peak in the relative energy spectrum at the frequency index equal to k and the value RES(f_k) is more than a threshold A_1, it is likely that there is a harmonic component at the frequency index k. The corresponding value RES(f_k) is assumed to be a measure of confidence of existence of the harmonic component.

Throughout a great number of experiments, it has been noted that, in most real music instruments, the several lowest harmonic components of the music notes are strong and can be extracted reliably through the second step of this method. Only a very low music note may have a very faint first harmonic component that cannot be extracted reliably. Based on these observations, some assumptions can be made for judging whether there is a pitch using the extracted harmonic components. In this method, some extra estimates are cancelled based on the following assumption:

If there is a pitch with a fundamental frequency of more than 82 Hz, either the lowest three harmonic components, or the lowest three odd harmonic components of this pitch, should all be present in the extracted harmonic components. If there is a pitch with a fundamental frequency that is lower than 82 Hz, four of the lowest six harmonic components should be present in the extracted harmonic components.

In two typical cases, the extra estimated pitches can be cancelled based on the above assumption. In the first case, the extra pitch estimation is caused by the noise peak in the preliminary pitch estimation. In the second case, the harmonic components of an extra estimated pitch are partly overlapped by the harmonic components of the true pitches. In this case, the non-overlapped harmonic components become important clues to check the existence of the extra estimated pitch. For example, if a polyphonic note contains two concurrent music notes C5 and G5, the fundamental frequency ratio of the two notes is nearly 2:3. Then, it is probable that there is extra pitch estimation on the C4, because the C4's second, fourth, sixth,…harmonic components are overlapped by the C5' first, second, third,… harmonic components, and the C4's third, sixth, ninth,… harmonic components are nearly overlapped by the G5's first, second, third,…harmonic components. However the C4's first, fifth, seventh harmonic components are not overlapped, so the extra C4 estimation can be easily cancelled by checking the existence of the first harmonic component based on the above assumption.

Through the last several steps, the extra incorrect estimation focuses on the pitches whose note intervals are 12 and 19 semitones higher than the true pitches. In this case, the fundamental frequencies of these extra estimated pitches are 2, 3or 4 times those of a true pitch and the harmonic components of each extra pitch are completely overlapped by a true pitch. For example, two of the estimated pitch candidates are the notes with fundamental frequencies f_1 and $3f_1$. Here, the difficulty is to determine if the note with the fundamental frequency $3f_1$ really occurs or, in fact, is an incorrect extra estimation caused by the overlapped frequency components of the lower music note. This difficult case is to be approached by the following observation. When a music note with the fundamental frequency f_1 is mixed with another note with the higher integer ratio fundamental frequency nf_1, then the corresponding harmonic spectral envelope often will not be smooth again and the spectral value of every n^{th} harmonic component becomes significantly larger than the neighbouring harmonic components. This can be measured by the Spectral Irregularity (SI) defined as follows,

$$SI(n) = \sum_{i=1}^{3} (RTFI(i \cdot n \cdot f_k) - (\frac{RTFI(i \cdot n \cdot f_k - 1) + RTFI(i \cdot \cdot f_k + 1)}{2})) \tag{7}$$

As indicated previously, if two of the estimated pitch candidates have the fundamental frequencies, f_1 and f_2 ($f_2 \approxeq nf_1$) and if the higher pitch does not occur, then the SI(n) is often smaller, according to the spectral smoothing principle. On the other hand, if the higher pitch does occur, then the overlapped harmonic components are often strengthened so that the $SI(n)$ has the larger value. So, in the proposed method, when the $SI(n)$ is smaller than a threshold, the overlapped higher pitch candidate is cancelled. The threshold is determined by experiments. In practical examples, most incorrect extra estimations caused by overlapping harmonic components are 2, 3, or 4 times the true pitches. Consequently, the proposed method only consider cases in which two pitch candidates have fundamental ratio at the 2,3 and 4.

4. RESULTS

In this task, our method performed third best according to the overall average accuracy. However, our method was also the most computationally efficient. The overall performance is reported in Table 2. On the one hand, the results from this task reflect the overall performance and efficiency differences between the submitted methods. On the other hand, it is difficult to find the specific reason that the method performs well or not, because the testing examples in the Mirex task are unknown. For some testing examples, our method has a relatively poor performance and produces many false positives. The reason for this probably is that there are many low-frequency notes in these examples. Generally speaking, the low-frequency notes have more complex spectral structures and cause the method to have more false positives.

Overall Average Accuracy	58.2%
Overall Average Precision	71.0%
Overall Average Recall	66.1%
Runtime (s)	271

Table 2. Overall results of Mirex 2007 multiple F0 frame-based estimation task for the submitted method

5. FUTURE WORK

In the proposed multiple pitch estimation method, temporal features are not exploited for the estimation. The method may be improved by utilizing the temporal features. The RTFI analysis can provide information about how the different frequency components of the analyzed signal evolve over time. This temporal information is useful for polyphonic pitch estimation. The harmonic components from the same instrument sound source often have some similar temporal features, such as a common onset time, amplitude modulation and frequency modulation. Harmonic relative frequency components with similar temporal features should be considered as a new pitch with more probability than the harmonic relative frequency components with different temporal features.

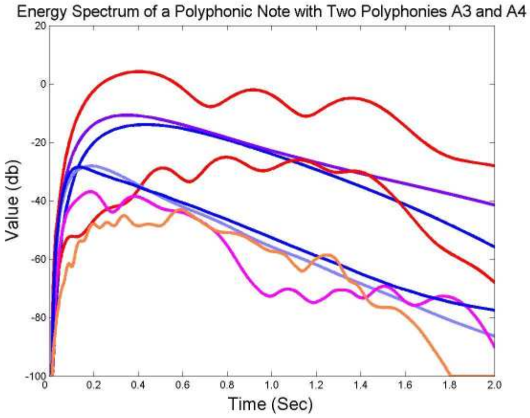

Figure 3. Energy Changes of Harmonic Components

For example, an analyzed polyphonic note consists of two polyphonies A3 and A4 The note A3 is played by piano and the note A4 is played by violin. It is very difficult to make a polyphonic estimation for this case, because the harmonic components of A4 are completely overlapped by the even harmonic components of A3. However, such a difficult case may be resolved by using the temporal feature. As shown in Figure 4, the blue lines denotes first four odd harmonic components of the note A3, and the red/magenta lines denotes the first four even harmonic components of the note A3. In Figure 4, it can be clearly seen that the energy spectrums of the first four even harmonic components have different temporal features than the first four odd harmonic components. This difference indicates that the even harmonic components probably are shared with another musical note A4 played by a different music instrument.

6. ACKNOWLEDGMENTS

We thank the MIREX organizers for their efforts to make this contest possible. In addition, this work is partially supported by the European project: EASAIER.

7. REFERENCES

[1] R.Zhou, *Feature Extraction of Musical Content for Automatic Music Transcription*, Ph.D. dissertation, Swiss Federal Institute of Technology, Lausanne, Oct, 2006. Downloadable on website http://library.epfl.ch/en/theses/?nr=3638.

[2] R.Zhou and M.Mattavelli, "*A new time-frequency representation for music signal analysis"* in Proc. International Conf. on Information Sciences, Signal Processing and its Applications, Sharijah, United Arab Emirates, Feb. 2007.