A REAL-TIME POLYPHONIC MUSIC TRANSCRIPTION SYSTEM

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 real-time polyphonic transcription system submitted to the multiple F0 note tracking task of 2008 Music Retrieval Evaluation eXchange (MIREX). The Resonator Time-Frequency Image (RTFI) is used as the basic time-frequency analysis tool. The evaluation of this system under the MIREX task is also discussed.

1. INTRODUCTION

Automatic music transcription is the process of converting an acoustical waveform into a musical notation (such as the traditional western music notation) by a computer program. It can be utilized to support a broad range of music applications. Beside the automatic music transcription itself, the possible applications include interactive music systems, low-bitrate compression coding for music signal and so on.

2. TIME-FREQUENCY ANALYSIS

2.1. Resonator Time-Frequency Image

A novel time-frequency representation, known as the Resonator Time-Frequency Image (RTFI), has been developed. Its main feature is that it selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis. This was chosen due to the flexibility with regards to time and frequency resolution, and the simplicity and computational efficiency of an implementation based on first order filters. The more detailed description of the RTFI can be found in [1] and [2].

2.2. 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. All 1080 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 music transcription system.

Let us define the RTFI average energy spectrum as follows,

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

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.

3. SYSTEM DESCRIPTION

3.1. System Overview

An automatic music transcription system has been constructed with the combination of a music onset

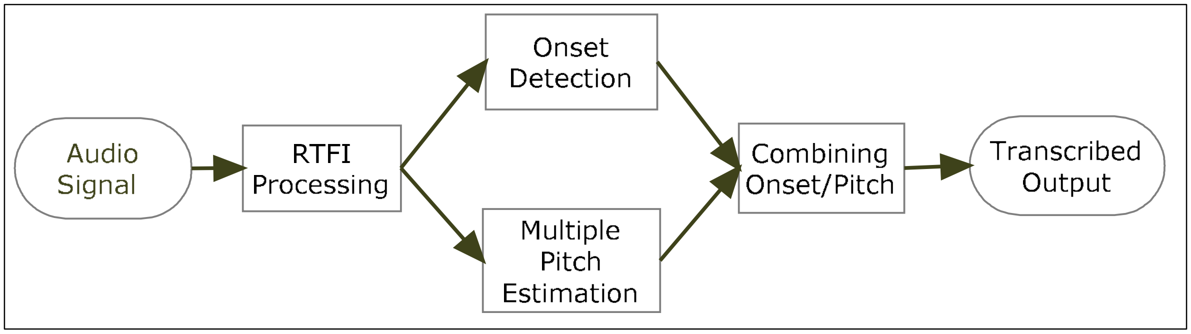

Figure 1 A real-time polyphonic music transcription system based on RTFI

detection algorithm and multiple pitch estimation method. The goal of the transcription system is to detect the music notes occurring, their onset times and note duration times. The sampling rate of input music signal is 44100 Hz. The onset detection algorithms are first used to separate the input real music signal into different segments according to the detected note onsets, and then pitches in each segment are estimated by the developed multiple pitch estimation method. Finally, every estimated pitch in a certain segment must be checked if the pitch begins from a current segment or from the previous segments. For a certain segment N, if a pitch A with fundamental frequency f is estimated; then if the estimated pitches in the previous segment N-1 do not contain the pitch A, the transcription system will consider that this pitch A is a new occurring pitch in the segment N. In another case when the estimated pitches in the previous segment N-1 also contain the pitch A, then this pitch A is considered to be a new occurring pitch only on the condition that the corresponding energy spectrum of the pitch A's first or second harmonic component has been obviously increased at the starting moment of the segment N. The note duration time is directly determined by how long the pitch continues to exit. Figure 1 shows the overview of the automatic music transcription system.

3.2. Onset Detection

Then RTFI average energy spectrum is further transformed into the pitch energy spectrum Y , smoothed pitch energy spectrum R, and difference pitch energy spectrum D according to the following equations:

$$R(k,\omega_m) = \frac{1}{5}\sum_{i=1}^{5} A(k, i\cdot\omega_m) \qquad (2)$$

$$S(k,\omega_m) = \frac{1}{25}\sum_{i=k-2}^{k+2}\sum_{m-2}^{m+2} R(k,\omega_m) \qquad (3)$$

$$D(k,\omega_m) = S(k,\omega_m) - S(k-n,\omega_m) \qquad (4)$$

where n is the difference order and N is the total number of frequency bins in the spectrum F .

In the proposed transcription system, an energy-based onset detection has been applied. A music signal is assumed into two parts - a transient part and a steady-state part. The difference pitch energy spectrum D can be used to track the transient information and generate an energy-based detection function as follows.

$$L(k,\omega_m) = H(D(k,\omega_m) - \theta_1), \quad \theta_1 > 0 \qquad (5)$$

$$DF(k) = mean(L(k,\omega_m)) \qquad (6)$$

where $H(x) = (x + |x|)/2$ is the half-wave rectifier function, and DF represents the energy-based detection function. The spectrum D is calculated with 3-order difference.

In the energy-based algorithm, firstly the difference pitch energy spectrum is limited by a threshold θ_1 so that only the energy-change values that exceed threshold θ_1 are considered to be possible transient clues; and then it is averaged across all frequency channels to generate the detection function. The detection function is further smoothed by a moving-average filter and a simple peak-picking operation is used to find the note onsets. In the peak-picking, another threshold θ_2 needs to be set and only the peaks having values greater than threshold θ_2 are considered as the possible onset candidates. In the final step, if there are two onset candidates and the position difference between them is smaller than or equal to 50ms, then only the onset candidate with the greater value will be kept.

3.3. Multiple Pitch Estimation

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

$$PES(f_k) = \frac{1}{L}\sum_{i=1}^{L} A(i\cdot f_k) \qquad (7)$$

$$RPES(f_k) = PES(f_k) - \frac{1}{N_2+1}\sum_{i=k-N_2/2}^{k+N_2/2} PES(f_i) \qquad (8)$$

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:

$$f_k = 440\cdot 2^{(k-690)/120} \qquad (9)$$

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):

$$PES(f_k) = \frac{1}{L}\sum_{i=1}^{L} RTFI(f_{k+A[i]}) \qquad (10)$$

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

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

Table 1 Deviation between approximation and ideal value

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.

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

$$RES(f_k) = A(f_k) - \sum_{i=k-N_1/2}^{k+N_1/2} A(f_i) \qquad (11)$$

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,

$$SI(n) = \sum_{i=1}^{3} (A(i \cdot n \cdot f_k) - (\frac{A(i \cdot n \cdot f_k - 1) + A(i \cdot \cdot f_k + 1)}{2})) \qquad (12)$$

As indicated previously, if two of the estimated pitch candidates have the fundamental frequencies, f_1 and f_2 ($f_2 \approx nf_1$) 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

We have submitted 3 transcription systems for evaluation. The results are summarized in Table 2. The 3 systems are very similar (as introduced above), but they have different numbers of filters. The numbers of their implementing filters are 1080, 360, 216 respectively. In the proposed music transcription systems, the main computational overheads depend on numbers of RTFI filters, so the 3 submitted transcription systems have different running time (as shown in Table 2). In this evaluation, our method performed third best according to the overall average accuracy, however our method is much more computationally efficient. Our method is almost 66 time faster than the second best one, and 5 time faster than the best one. In addition, our method performed best on the piano subset. As the parameters of our transcription systems were tuned only by a piano and guitar dataset, the overall performance can be improved if the parameters of the method are tuned by different datasets.

	ZR1	ZR2	ZR3
Overall Average F-measure	0.518	0.520	0.530
Overall Average Recall	0.602	0.604	0.600
Overall Average Precision	0.466	0.467	0.486
Average F-measure on Piano	0.757	0.754	0.743
Average Recall on Piano	0.777	0.775	0.744
Average Precision on Piano	0.738	0.734	0.743
Running time (second)	2700	1451	871

Table 2 Result of the polyphonic music transcription system submitted to MIREX 2008

5. 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.

6. 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.