MUSIC ONSET DETECTION COMBINING ENERGY-BASED AND PITCH-BASED APPROACHES

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

This paper describes a new music onset detection method submitted to the Mirex 2007 onset detection task. The method is based on the combination of energy-based and pitch-based detection. The Resonator Time Frequency Image (RTFI) is utilized for a basic time-frequency analysis tool.

Keywords: MIREX, onset detection, RTFI

1. INTRODUCTION

Music note onsets may be classified as "soft" or "hard". A hard onset is accompanied by a sudden change in energy, whereas a soft onset shows a more gradual change. With the appropriate time-frequency representation, the hard onsets can be easily detected by energy-based algorithms. However, the detection of a soft onset is a much more difficult task because music signals often contain noise and oscillations associated with frequency and amplitude modulation. The energy-change in the oscillation often surpasses the energy-change of a soft onset and this fact makes it very difficult to distinguish true onsets from other changes if relying only on the energy-change cue.

This paper describes a new method that makes best use of both energy-change and pitch-change information. As shown in Figure 1, the method consists of three main stages: time-frequency processing, onset type classification and detection algorithms. This paper briefly describes the basic idea behind the method and reports on its analysis in the Music Information Retrieval Evaluation eXchange (MIREX). More detailed information about the time-frequency processing and detection algorithms are reported in [1].

2. TIME-FREQUENCY PROCESSING

2.1. Resonator Time-Frequency Image (RTFI)

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. The more detailed description of the RTFI can be found in [1] and [2].

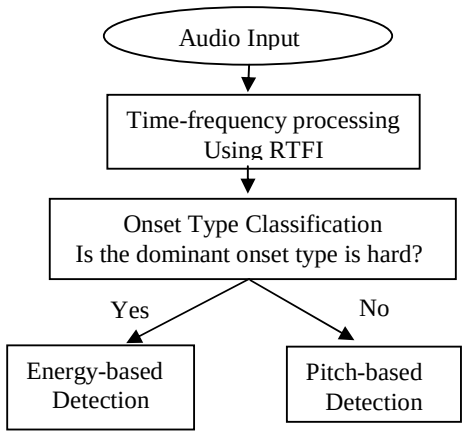

Figure1. System overview of the proposed method

2.2. Time-Frequency Processing

The monaural music signal is used as the input signal at a sampling rate of 44.1 kHz. The method utilized RTFI as the basic tool for time-frequency processing. The center frequencies of discrete RTFI are according to logarithmic scale and the resolution is selected as constant-Q. 10 filters are used to cover the frequency band of one semitone and there is a total of 960 filters in the analyzed frequency range, which extends from 46 Hz to 6.6 kHz. Consequently, the average RTFI energy spectrum can be obtained and expressed as follows,

$$ARTFI(k,\omega_m) = db(\frac{1}{M}\sum_{n=(k-1)M+1}^{kM}\left|RTFI(n,\omega_m)\right|^2) \quad (1)$$

where the M is an integer, k is the index of time frame, *db()* converts the value to decibels and the ratio of M to sampling rate is the duration time of the frame in the average process. In this paper, the average RTFI energy spectrum is calculated in unit of 0.01 second and it is used to track the time-frequency character of music signal.

To remove the low-frequency noise, the RTFI energy spectrum is first transformed into the adjusted energy spectrum (AES) according to the Robinson and Dadson equal-loudness contours, which have been standardized in the international standard ISO-226. In order to

simplify the transformation, only an equal-loudness contour corresponding to 70db is used to adjust the RTFI energy spectrum. The standard provides equal-loudness contour limited to 29 frequency bins. Then, this contour is used to obtain the equal-loudness contour of 960 frequency bins by cubic spline interpolation in the logarithmic frequency scale. Let us define this equal-loudness contour as $Eq(\omega_m)$ in db. Then, the adjusted energy spectrum (AES) can be expressed as follows:

$$AES(k,\omega_m) = ARTFI_E(k,\omega_m) - Eq(\omega_m) \quad (2)$$

In the adjusted energy spectrum, one can select a threshold value of the energy spectrum below which it will be considered as a noise spectrum. Then the adjusted energy spectrum is further transformed into the pitch energy spectrum *R*, smoothed pitch energy spectrum *S*, difference pitch energy spectrum *D* and the normal pitch energy spectrum *F* according to the following equations:

$$R(k,\omega_m) = \frac{1}{5}\sum_{i=1}^{5} AES\ (k, i\cdot\omega_m) \quad (3)$$

$$S(k,\omega_m) = \frac{1}{25}\sum_{i=k-2}^{k+2}\sum_{m-2}^{m+2} R(k,\omega_m) \quad (4)$$

$$D(k,\omega_m) = S(k,\omega_m) - S(k-n,\omega_m) \quad (5)$$

$$F(k,\omega_m) = S(k,\omega_m) - \max((S(k,\omega_m))_{m=1:N}) \quad (6)$$

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

In practical cases, instead of using equation (3), the spectrum *R* can be easily calculated in the logarithm scale by the following approximation:

$$R(k,\omega_m) \approx \frac{1}{5}\sum_{i=1}^{5} Y(k,\omega_{m+A[i]}) \quad (7)$$

$$A[5] = [0,120,190,240,279] \quad (8)$$

i	1	2	3	4	5
$\frac{\omega_{m+A[i]}}{i\cdot\omega_m}$	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.

The difference pitch energy spectrum *D* makes the energy change more obvious, and the normal pitch energy spectrum *F* makes pitch change more clearly.

3. DETECTION ALGORITHMS

3.1. Onset Type Classification

A simple technique is used to classify the dominant onset type of the analyzed input file. The measure of the onset 'hardness' is defined as follows:

$$Q(k) = mean(H(D(k,\omega_m))) \quad (9)$$

$$HM = mean(Q(k)) \quad (10)$$

where $H(x) = (x + |x|)/2$ is the half-wave rectifier function, and spectrum *D* is calculated with first order difference. The hardness measure *HM* is used to classify the dominant onset type. If the *HM* of the analyzed input file is more than a threshold, the onset type of this input is considered as hard, otherwise it is considered to be soft.

For the input with hard onsets, the energy-based algorithm is used to find onsets. Conversely, the pitch-based algorithm is utilized with soft onsets.

3.2. Energy-based detection

A music signal is assumed to be comprised of 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.

$$L(k,\omega_m) = H(D(k,\omega_m) - \theta_1), \quad \theta_1 > 0 \quad (11)$$

$$DF(k) = mean(L(k,\omega_m)) \quad (12)$$

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. Pitch-based detection

Generally speaking, energy-based detection methods are not good at detecting soft onsets. Consequently, a pitch-based algorithm has been developed. In the proposed pitch-based detection algorithm, the music signal is first divided into transient and stable parts by the stable pitch

cue, and then the onset is located in the transient part by energy-change. As the output of RTFI time-frequency processing, the spectrum *D* and the spectrum *F* are used together as the input for this detection algorithm. The algorithm can be separated into two steps:

1) Searching the possible note onsets with the approximate fundamental frequency ω_{m1}.

2) Combining the detected onset candidates across all of the frequency channels and generating the final result for onset detection.

In the first step, the algorithm searches possible note onsets in every frequency channel. It is emphasized that, when searching in a certain frequency channel with frequency ω_{m1}, the detection algorithm tries to find only the onset where the new occurred pitch rightly has an approximate fundamental frequency ω_{m1} .

If a pitch with a fundamental ω_{m1} occurs in a certain time segment, then there is often a peak line in this time segment around the frequency ω_{m1} in the spectrum *F* and its value is nearly equal to 0db and relatively larger than the other frequency bins. This fact has been observed in our experiments.

When searching for onsets in a certain frequency channel with frequency ω_{m1}, the detection algorithm first tries to find the "stable time segment *T* ", which corresponds to the steady-state part of a music note. Let us suppose the time segment $T[k_1,k_2]$ represents a time duration from k_1 to k_2 in units of 10ms. Given a time-frequency spectrum $F(k, \omega_m)$, if a time segment $T[k_1,k_2]$ meets with the following three conditions, the time segment *T* is assumed to be stable.

$$(F(k,\omega_m))_{m=m1,k=k1:k2} > \alpha_1 \tag{13}$$

$$\max((F(k,\omega_m))_{m=m1,k=k1:k2}) > \alpha_2 \tag{14}$$

$Sum(\omega_m)$ has a spectral peak at the frequency ω_{m1},

$$Sum\ (\omega_m) = \sum_{k=k_1}^{k_2} F\ (k\ ,\omega_m\ ) \tag{15}$$

For each stable time segment $T[k_1,k_2]$, the algorithm looks backward from beginning of the stable time segment *T* and locates the onset time in 300ms window by searching salient energy-increasing in the duration $[k_1\text{-}300,\ k_1]$. The salient energy-increasing is defined by peak-picking in the different pitch spectrum *D* in the duration $[k_1\text{-}300ms,\ k_1]$ at the frequency ω_{m1}. The threshold α_3 of the peak-picking process is the third important parameter for this algorithm.

After all frequency channels have been searched, the pitch onset candidates can be found and expressed as follows:

$$Onset\_C(k,\omega_m) \geq 0\ ,\ m=1,\ 2,\ 3,\ \ldots N, \tag{16}$$

where *k* denotes the time frame and *N* denotes the total num of the pitch channels. If $Onset\_C(k,\ \omega_m)=0$, no onset exists in the k_{th} time-frame and m_{th} frequency channel. If the $Onset\_C(k,\ \omega_m)>0$, there is an onset candidate in the k_{th} time-frame and m_{th} frequency channel, and the value of $Onset\_C(k,\ \omega_m)$ is equal to the value of spectrum $D((k,\ \omega_m)$.

Finally the detection algorithm combines the pitch onset candidates across all the frequency channels to get the final onset. If two onset candidates are neighbors in a 0.05 second time window, then only the onset candidate with the greater value will be kept.

4. RESULTS

According to the overall performance, our method outperforms all other techniques which were evaluated in this task (reported in Table 2). In particular, our method performed best on the overall average F-measure, which was the primary criterion for evaluation. Different methods can perform significantly better for different classes. Our method performs better than the other methods for the classes: solo drum, solo brass and solo wind. On the other hand, our method need more running time than other methods. Stowell's method (stowell_cd) and Lacoste's method are very computationally efficient.

For the class of bars and bells, our method performs relatively poorer according to the average F-measure, but with a precision of 100%. At least for this class, the parameter values of the energy-based detection are set too small. It is expected that the performance can be improved by selecting better parameter values.

In the MIREX 2005~2007 onset detection contests, most of the submitted methods are energy-based. The results suggest that a common difficulty exists in the onset detection of the classes: solo brass, solo wind, solo sustained string and solo singing voice. These classes usually contain a large number of soft onsets. Energy-based approaches are based on the assumption that there are relatively more salient energy changes at the onset times than in the steady-state parts. In case of soft onsets, the assumption can not stand. The significant energy changes in the steady-state parts can mislead energy-based approaches and cause many false positives. According to our previous experiments, the pitch-based detection can clearly outperform the energy-based detection for the detection of soft onsets.

Overall Average F-measure	80.8%
Overall Average Precision	85.7%
Overall Average Recall	78.2%
Total Correct	7225
Total False Positives	1186
Total False Negatives	2130
Total Merged	189
Total Doubled	49
Runtime (s)	1399

Table 2: Overall results of Mirex 2007 onset detection task for the submitted method

For the solo brass and solo wind, our method outperforms the second best methods by about 8% and 9% respectively. Such performances can be contributed to the combination of the pitch-based detection. For the single sustained string class, the method is the second best one and Lee's method is better. This can be explained that Lee's method is not only energy-based but also combines a phase-based detection, which uses the stable pitch cues indirectly.

For the solo singing voice class, our method performs not well. The reason is that the method is developed on the music datasets played by musical instruments. In the steady-state parts, the pitch variation of instrumental music is minor, but the singing voice's pitch variation is relatively larger. The method need be improved to achieve a better performance on the onset detection of singing voice. Robel's method performs best for the onset detection of the singing voice class.

5. FUTURE WORK

In this contest, the method is simulated on Matlab. The RTFI is implemented by Matlab mex function, the other parts are implemented by Matlab language. The running time need 1.6 times than real-time. In the future work, the speed of the method can be optimized by using the multi-resolution faster RTFI, which has been developed in [1].

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.

[3] W.L, Y.Shiu and C.J.Kuo, "*Musical onset detection with linear prediction and joint feature* " MIREX 2007 audio onset detection contest: http://www.music-ir.org/mirex2007/abs/OD_lee.pdf

[4] D.Stowell, M.Plumbery, "*Adaptive whitening preprocessing applied to onset detectors, MIREX 2007* " MIREX 2007 audio onset detection contest: http://www.musicir.org/mirex2007/abs/OD_stowell.pdf

[5] A.Lacoste, " *Turbo convolution 2000*" MIREX 2007 audio onset detection contest: http://www.music-ir.org/mirex2007/abs/OD_lacoste.pdf

[6] A.Robel, " *Onset detection in polyphonic signals by means of transient peak classification"* MIREX 2007 audio onset detection contest: http://www.music-ir.org/mirex2007/abs/OD_roebel.pdf