MIREX 2006: Spectral-flux based Musical Onset Detection

Yunfeng Du
Institute of Acoustics,
Chinese Academy of Sciences
ydu@hccl.ioa.ac.cn

Ming Li
Institute of Acoustics,
Chinese Academy of Sciences
mli@hccl.ioa.ac.cn

Jian Liu
Institute of Acoustics,
Chinese Academy of Sciences
jliu@hccl.ioa.ac.cn

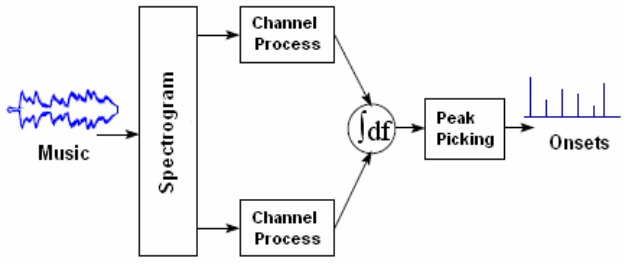
Abstract

This abstract describes a submission of the ThinkIT Speech Lab at Institute of Acoustics, Chinese Academy of Sciences to the onset detection contest of the Music Information Retrieval Evaluation eXchange (MIREX) 2006. This submission presents an algorithm using spectral flux to detect musical onsets. Firstly, a detection function is generated via spectral flux. Then, a peak picking procedure is applied on this function to extract the onset points. Finally, the evaluation result of the submitted algorithm is presented with discussion and analysis.

Keywords: MIREX, onset detection, spectral flux

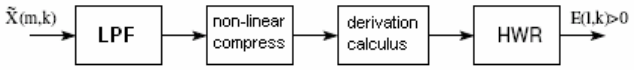
1. System Overview

The submitted onset detection algorithm mainly uses the spectral flux [1] to generate detection function for measuring the musical onset regions, followed by a peak picking procedure to extract the onset points in the audio stream.



Onset detection system overview

The detection function is generated via the flux of spectral energy, the main steps include (1) FFTs based on time sequence to build a spectrogram for the input audio, (2) low pass filtering, μ-law compression, Canny-operator based differentiating and half-wave-rectification for each frequency channel of the spectrogram, (3) summating all frequency channels together to form the detection function. The information of spectral phase is ignored due to the tiny contribution to the overall performance.



Processing performed at each frequency channel

The peak picking procedure is mainly achieved by thresholding, the main steps include (1) picking up all local maxima, (2) computing threshold based on the standard deviation of the detection function, (3) picking up all the local maxima which are beyond the threshold, (4) merging the onset points which are too close to each other. All the surviving onset points make up of the final result.

2. Description of Algorithm

The components of the presented algorithm are described detailedly in this section. Firstly, the components used in generating the detection function are introduced, and then all the arts utilized in the peak picking procedure to extract onset points are presented. The input audio stream is down-sampled into 16 kHz with uniform format of 16 bits, mono channel.

2.1 Generating detection function

Firstly, a computation of spectrogram is achieved by applying FFT on each frame of the audio stream. The frame length is 16ms, and the FFT size is set as 512 which is a double size of the frame length to promote the spectral resolution. Further processing is applied on each frequency channel of the spectrogram, which is represented as a function of spectral energy at a certain FFT bin with respect to time. After each frequency channel being processed, all channels are integrated together to form the detection function. Below are the detailed processes applied in each frequency channel.

2.1.1 Low pass filtering

A low pass filtering operation is first applied on each frequency channel to extract the spectral energy envelope.

This procedure is achieved by convolving the frequency channel function with a Half-Hanning window [2, 4] which has a low-pass characteristic. The length of the Half-Hanning window is set as 50ms in our algorithm.

2.1.2 μ-law compression

After low pass filtering, a μ-law compression [3] is applied on each filtered channel to achieve a non-linear compression effect. This procedure together with the following differentiating procedure is a psycho-acoustic relevant process that can catch a tiny but perceptible spectral energy change more precisely. The compression factor μ which determines the degree of compression is set as 100 in our algorithm.

2.1.3 Differentiating

After μ-law compression, a differentiating procedure is applied on each filtered and compressed channel to transform the sudden rises of spectral energy into narrow peaks. The differentiating is achieved by using Canny-operator which is widely used in image processing [4]. The σ, which controls the shape of the operator, is set as 1 in our algorithm.

2.1.4 Half-wave rectification (HWR)

In a similar way, after differentiating, there will be negative peaks exhibited in each channel when the spectral energy drops, marked as an offset. Since we are only interested in onsets, a half-wave rectification (HWR) is applied to only preserve the positive peaks in each channel.

2.2 Extracting onset points

When the detection function is generated, a peak picking procedure is applied on this function to search out all the onset points. Below are the detailed steps.

2.2.1 Local maxima searching

Local peaks' positions and heights are detected in the detection function with a running window method: local maxima are detected at the indexes whose values are higher than those of their neighbours within 25ms.

2.2.2 Thresholding

The threshold is computed as the product of an adjustable coefficient and the standard deviation of the detection function, whose value is computed only based on the non-zero values in the function. Then all the local maxima beyond the threshold are picked as onset candidates.

2.2.3 Candidates merging

The minimum duration between every two reasonable onsets is set as 100ms. Therefore, all the candidate points, which are too close to each other, are merged together by preserving the one with greater magnitude and deleting the weaker one. All the surviving onset points make up of the final result.

3. Implementation

The implementation of this algorithm is achieved by C++ and built in Win32 environment with Intel C++ Complier 9.0. The implementation consists of a front-end utilizing the program of "sox" [5] to achieve resample for the input audio steam, whose format is required to be the Windows PCM with WAV header.

4. Evaluation Result

5. Acknowledgement

This work is supported by Chinese National 973 program (2004CB318106) and National Natural Science Foundation of China (10574140, 60535030). Many thanks to the IMIRSEL for running the evaluation.

References

[1] M. Alonso, "Extracting Note Onsets from Musical Recordings", *Proc. IEEE International Conference on Multimedia and Expo*, 2005

[2] E. Scheirer, "Tempo and Beat Analysis of Acoustic Music Signals," J. Acoust. Soc. Am., vol. 103, no. 1, pp. 588-601, Jan. 1998

[3] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the Meter of Acoustic Musical Signals", *IEEE Trans. Audio, Speech, and Language processing*, vol.14, no.1, Jan. 2006

[4] Lie Lu, Hong-Jiang Zhang. "Automated Extraction of Music Snippets ", *Proc. ACM Multimedia 03*, pp.140-147, Berkeley CA, Nov. 2-8, 2003

[5] http://sox.sourceforge.net/