An Energy-based and Pitch-based Approach to Audio Onset Detection

Hui Li Tan, Yongwei Zhu, Lekha Chaisorn
Institute for Infocomm Research, A*STAR,
Singapore
{hltan, ywzhu, clekha}@i2r.a-star.edu.sg

ABSTRACT

This paper presents an approach for audio onset detection submitted to MIREX 2009. The technique utilizes information on the general characteristics of the notes for onset categorization, as well as integrates energy-based and pitch-based detection results.

1. INTRODUCTION

Music note onsets are signified by the presence of changes in energy and/or pitch content of the audio either abruptly or gradually. The characteristics of the changes are due to the types of instruments and the performance techniques that produce the notes. Hence, audio onset detection can be challenging since the change characteristics of note onsets can vary across or within music pieces.

In this paper, an approach for audio onset detection is presented. The technique utilizes information on the general characteristics of the notes for onset categorization, as well as integrates energy-based and pitch-based detection results. The onset categorization, audio energy-based and pitch-based processing, and onsets integration steps will be covered in Sections 2, 3 and 4 respectively. Results and findings on a self-acquired dataset as well as the MIREX audio onset detection dataset will be presented in Section 5. Eventually, we conclude with some of our future research directions in Section 6.

2. ONSET CATEGORIZATION

Due to the inherent variable nature of the beginning of audio onsets resulting from the different types of instruments, no simple algorithm will be optimal for general onset detection. Hence an audio signal is first roughly categorized into three main groups to facilitate further processing. The three groups are:
(1) Un-pitched Onsets (hard onsets) which consist drums and other percussion instrument onsets;
(2) Pitched Percussive Onsets (hard onsets) which consist bars & bells, brass, plucked strings and sustained strings (struck strings e.g. piano) onsets;
(3) Pitched Non-percussive Onsets (soft onsets) which consist sustained strings (bowed strings e.g. violin) and winds onsets.

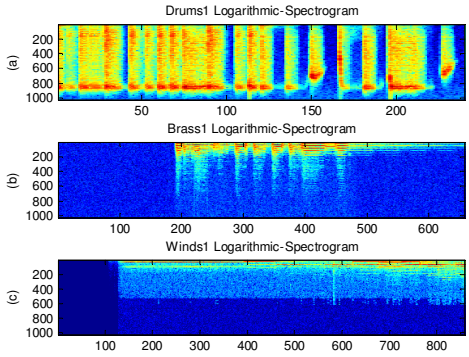

Figure 1. Logarithmic-Spectrogram of a (a) drum excerpt. Un-pitched onsets are typically characterized by wide band and transient nature. (b) brass excerpt. Pitched percussive onsets have significant energy changes corresponding to pitch changes. (c) winds excerpt. Pitched non-percussive onsets demonstrate pitch changes with only gradual energy attacks.

The categorization is done by percussiveness estimation and pitch tuning estimation which are elaborated in sub-sections 2.1 and 2.2 respectively.

2.1 Percussiveness Estimation

In percussive estimation, we aim to detect the average bandwidth of abrupt energy changes in the audio signal. In this process, energy changes in both band limited (1K-5KHz) and whole band are utilized. A potential percussive onset is claimed when the band limited energy change with high magnitude is present. The average bandwidth of all such potential percussive onsets are computed, and used as an indicator for the presence of highly percussive onsets in the signal. The estimated percussiveness can roughly distinguish highly percussive onsets to non-percussive onsets, such as from the brass to the winds.

2.2 Pitch Tuning Estimation

In pitch tuning estimation, we aim to detect the presence of tuned pitch in the music excerpt. The tuning detection is conducted on the CQT spectrogram of the audio signal [1], with a pitch resolution of 10 points per semitone being used. After peak picking and grouping, the energy concentration of the peaks across a 10-bin histogram is investigated. The bin number indicates the deviation of the pitch content of the excerpt from a standard tuning reference pitch, e.g. 440Hz for middle A. Then, both time duration and energy percentage of the tuned pitch in the

audio signal are computed and thresholds are used to estimate the presence of pitch in the signal.

2.3 Heuristic Rules for Categorization

Heuristic rules were used for the categorization of onsets. The estimated percussiveness and pitch tuning were represented by probability values, and empirical thresholds were imposed to put the audio piece into one of the three categories.

3. AUDIO PROCESSING

Motivated by [2], we employ energy-based and pitch-based approaches for audio processing. These are elaborated in sub-sections 3.1 and 3.2.

3.1 Energy-based Processing

Energy percussiveness is an essential feature and hence, we use our enhanced percussiveness measure, *D(i)* as:

$$D(i) = \sum_{j \in J} \begin{matrix} 0 & if \quad d_i(j) < T \\ 1 & if \quad d_i(j) \geq T \end{matrix} \qquad (1)$$

$$\text{where } d_i(j) = \log_2 \left(\frac{\left| STFT_x^w(j,i) \right|}{\left| STFT_x^w(j,i-1) \right|} \right),$$

and $STFT_x^w(j,i)$, is the Short-Time Fourier Transform computed with hop x, window w, and with the hamming window. $i \in [1, I]$ represents the time index while j represents the band index and J defines the spectral range over which the distance is evaluated. The threshold T is experimentally set.

For the un-pitched onsets, energy-based detection performance is optimized by considering the full frequency band from 0-22050Hz. In contrast, for the pitched onsets, energy-based detection performance is optimized by considering the range of musical instruments, that is $J \in$ [1KHz, 5KHz].

An adaptive threshold [3] was then applied on the measure. Time frames with sufficient counts above the threshold were then declared as potential onsets.

3.2 Pitch-based Processing

Let C_{STFT} be the chromagram of a music excerpt. The strength of the base and dominant harmonics pitch class pairs are then computed by summing the magnitude of the base pitch classes with its respective dominant harmonics pitch classes, given by

$$S(k_1,i) = C_{STFT}(k_1,i) + C_{STFT}(k_8,i), \qquad (2)$$

for $\forall i$, and where $k_8 = (k_1+6) mod 12+1$; k_1, $k_8 \in [1, 12]$. The strongest base pitch class and dominant harmonics pitch class pair for each time frame i is then the optimizer for $\max_{k_1}(S(k_1,i))$.

Changes in the strongest pitch pair suggest changes in pitch, and mark potential onsets. In contrast, stability in the strongest pitch pair suggests stable pitch content corresponding to an onset. Hence adjacent time frames with the same pitch content are grouped together in a cluster. Refer to Figure 2 for the strongest pitch pair extraction for a short plucked strings excerpt, illustrating clusters of strongest pitch pair corresponding to onsets.

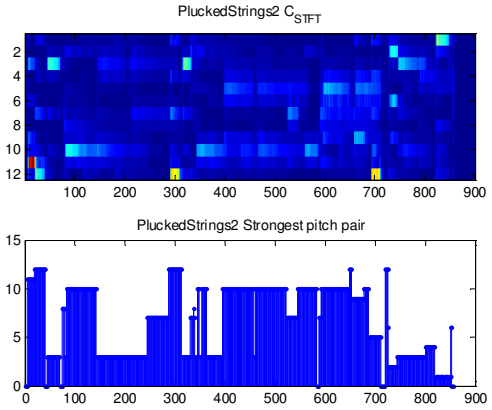

Figure 2. For a short plucked strings excerpt, (Top) Chromagram. (Bottom) Strongest pitch pair.

Fluctuation of the strongest pitch pair due to energy leakage could result in over-fragmentation. Therefore, the clusters were further analyzed through the chromagram extracted from the note partials components extracted according to [1], C_{CQT}. The tuning pitch determination and note partial components extraction algorithms were used to extract precisely the note partials from the signal, while allowing for as much pitch bending/vibration as possible. We then compute the pitch profile, P, for each cluster by:

$$P(k,c) = \sum_{i=c_s}^{c_n} C_{CQT}(k,i), \qquad (3)$$

for $k \in [1, 12]$, and where $c \in [1,$ total number of clusters$]$, c_s and c_n denotes the starting and ending frames of cluster c respectively. Pitch profile features were extracted and used in determining if two adjacent clusters should be merged. Eventual clusters indicate regions pertaining to the individual onsets.

4. ONSETS INTEGRATION

After onset categorization, a music excerpt is roughly put into one of the three categories. For the un-pitched category, onset detection is based only on energy processing. For the pitched percussive and pitched non-percussive onset detection, both energy and pitch information are used.

4.1 Un-pitched Onset Detection

The detection is based on energy change information. Changes detected in the band limited and whole band energy are integrated with time alignment and double thresholding.

4.2 Pitched Percussive and Non-percussive Onset Detection

In these categories, both pitch change and energy change detections are used in onset detection. The pitch-based detections are aligned to the energy-based detection, as the energy-based detection generally provides better time precision.

For the percussive category, a potential onset is declared as an onset if its energy change is significant, regardless of the presence of pitch change. In contrast, for the non-percussive category, a lower weight is put to the energy-based detections and an onset can be claimed when there is pitch change without corresponding energy change.

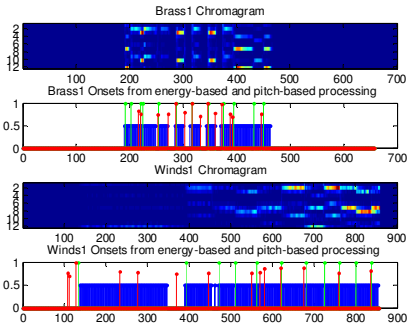

Figure 3. Onsets integration. For brass excerpt, (a) Chromagram. (b) Energy-based detections (Red) synchronized with pitch-based detections (Blue) often match true onsets (Green). True onsets may also match energy changes without pitch change. For a winds excerpt, (c) Chromagram. (d) Energy-based detections (Red) synchronized with pitch-based detections (Blue) often match true onsets (Green). True onsets may also pitch changes without significant energy change.

5. RESULTS

Before submission, preliminary testing of the energy-based detection and pitch-based detection performances were conducted on a self-acquired dataset. The self-acquired dataset comprises 7 categories in Table 1. There are 5-10 excerpts for each category and each excerpt is 5-10 second long.

Table 1 illustrates the detection results from the energy-based processing and pitch-based processing. The detection performances for the brass and winds improved, indicating that pitch-based information does aid in audio onset detection. Nonetheless, the detection performance for the plucked strings, sustained strings and vocals dropped, due to a drop in precision, possibly led by over-fragmentation in the pitch-based processing. The drop in recall could also be due to missing chord without changes in the strongest pitch pair. As illustrated in Figure 2, the onset at time frame 217 was not detected as the onset was due to a change from monophonic note to major chord note, with its strongest pitch pair remaining unchanged.

	Energy-based			Pitch-based		
	P	**R**	**F**	**P**	**R**	**F**
Drums	0.87	0.90	0.88	0.71	0.64	0.67
Brass	0.66	0.68	0.67	0.90	0.87	0.89
Plucked Strings	0.64	0.90	0.75	0.66	0.69	0.67
Sustained Strings	0.68	0.61	0.65	0.61	0.59	0.60
Vocals	0.68	0.54	0.60	0.35	0.75	0.48
Winds	0.43	0.36	0.40	0.75	0.59	0.66
Mixes	0.93	0.80	0.86	0.71	0.63	0.67

Table 1. Detection results from the energy-based and pitch-based processing. (P – Average Precision, R - Average Recall, F – Average F-measure).

For the MIREX audio onset detection, we submitted 5 algorithms, with varying extent of pitch-based information being incorporated. The first submission, TZC1, which has the most pitch-based information being incorporated performed overall best compared to our four other submissions, indicating the utility of the pitch-based detection.

The MIREX dataset is subdivided into 9 classes and the performance of the algorithms are assessed on each class separately, as well as the whole dataset. The result for one of our submission is shown in Table 2. As shown, independent of the dataset, the system performed well for the solo brass and solo winds. For the other categories, our proposed system performed moderately well compared to the best results, considering that there exists much room for parameters adjustments and better clustering decisions. Relatively preliminary heuristic rules and features were used for onset categorization and in the merging decision of clusters, and this is an area for further analysis. For the solo drum, our algorithm has high recall compared to other submissions. Nonetheless, the high doubled detection resulted in a lower precision. This is another issue for further analysis.

Class	**P**	**R**	**F**	**Best F**
Complex	0.729	0.593	0.643	0.748
Poly Pitched	0.859	0.831	0.831	0.915
Solo Bars and Bells	0.879	0.973	0.917	0.994
Solo Brass	0.734	0.766	0.746	0.799
Solo Drum	0.846	0.904	0.861	0.907
Solo Plucked Strings	0.674	0.831	0.725	0.901
Solo Singing Voice	0.266	0.492	0.333	0.395
Solo Sustained Strings	0.677	0.443	0.522	0.64
Solo Winds	0.744	0.765	0.753	0.753
Total	**0.757**	**0.77**	**0.744**	**0.796**

Table 2. Detection results from the MIREX dataset. (P – Average Precision, R - Average Recall, F – Average F-measure, Best F – Best F-Measure).

6. CONCLUSION AND FUTURE WORK

Relatively preliminary heuristic rules have been used for onset categorization and this could be improved for more accurate classification, so that the mixed instrument case could be better handled. Moreover, attention would be given to investigate and incorporate better pitch profile features and high-level information involving chord analysis for better clustering decisions. This will minimize false detection due to over-fragmentation, and hence help in improving the robustness of the pitch-based detection performance.

The solo vocals, one of the more challenging category, achieved a best F-measure of only 0.395. It would be interesting to explore how speech processing techniques could be encompassed to improve its detection performance.

The audio onset detection is the fundamental starting block for many high level music information retrieval tasks such as rhythm analysis [4, 5], audio chord detection, audio beat tracking [6] etc. This will greatly aid the development of our future music search and recommendation system.

7. ACKNOWLEDGEMENT

We would like to express our gratitude to the organizers of MIREX 2009 for the enormous effort in organizing the contest. Invaluable research pointers have been gained, and these helped shape our future research directions.

8. REFERENCES

[1] Y. Zhu and M. Kankanhalli, "Precise Pitch Profile Feature Extraction from Musical Audio for Key Detection", *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp.575-584, June 2006.

[2] R. Zhou, JD. Reiss, "Music onset detection combining energy-based and pitch-based approaches", *elec.qmul.ac.uk.*

[3] JP. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. Sandler, "A tutorial on onset detection in music signals", *IEEE Transactions on Speech and Audio Processing, 2005,* vol. 13, no. 5, pp. 1035-1047.

[4] Y. Zhu, HL. Tan, S. Rahardja, "Drum loop pattern extraction from polyphonic music audio", *IEEE Intl Conference on Multimedia and Expo, 2009, ICME 2009*, pp. 482-485, June 2009.

[5] HL. Tan, Y. Zhu, S. Rahardja, L. Chaisorn, "Rhythm analysis for personal and social music applications using drum loop patterns", *IEEE Intl Conference on Multimedia and Expo, 2009, ICME 2009*, pp. 1672-1675, June 2009.

[6] A. Klapuri, M. Davy 2006. Signal processing methods for music transcription. Hainsworth, S., Beat tracking and musical metre analysis. pp. 101-127, Springer.