Recognition of Note Onsets in Digital Music Using Semitone Bands

Antonio Pertusa[1], Anssi Klapuri[2], and José M. Iñesta[1]

[1] Departamento de Lenguajes y Sistemas Informáticos,
Universidad de Alicante, Spain
{pertusa, inesta}@dlsi.ua.es
[2] Signal Processing Laboratory,
Tampere University of Technology, Finland
klap@cs.tut.fi

Abstract. A simple note onset detection system for music is presented in this work. To detect onsets, a 1/12 octave filterbank is simulated in the frequency domain and the band derivatives in time are considered. The first harmonics of a tuned instrument are close to the center frequency of these bands and, in most instruments, these harmonics are those with the highest amplitudes. The goal of this work is to make a musically motivated system which is sensitive on onsets in music but robust against the spectrum variations that occur at times that do not represent onsets. Therefore, the system tries to find semitone variations, which correspond to note onsets. Promising results are presented for this real time onset detection system.

1 Introduction

Onset detection refers to the detection of the beginnings of discrete events in an audio signal. It is an essential component of many systems such as rhythm tracking and transcription schemes. There have been many different approaches for onset detection, but it still remains an open problem.

For detecting the beginnings of the notes in a musical signal the presented system analyses the spectrum information across 1/12 octave (one semitone) bands and compute their relative differences in time to obtain a detection function. Finally, the peaks in this function that are over a threshold are considered as onsets.

There are several onset detection systems that apply a pre-processing stage by separating the signal into multiple frequency bands. In an onset detector introduced by Klapuri [1], a perceptually motivated filter-bank is used, dividing the signal into eight bands. Goto [2] slices the spectrogram into spectrum strips [3]. Scheirer [4] uses a six band filter-bank and Duxbury *et al* [5] utilizes a filterbank to separate the signal into five bands.

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 869–879, 2005.

In the well-tempered scale, the one used in western music, the first harmonics[1] of the tuned instrument notes are close to the center frequencies of the 1/12 octave bands. In most instruments these first harmonics are those with the highest amplitudes.

It is not our aim to use a perceptually motivated approach. Instead, a musically motivated filter-bank is utilized. In music, notes are separated by semitones, so it makes sense to use a semitone filterbank to detect their onsets. By using semitone bands the effect of subtle spectrum variations produced during the sustain and release stage of a note is minimized. While a note is sounding, those variations mainly occur close to the center frequencies of the 1/12 octave bands. This means that the output band values for a note will remain similar after its attack, avoiding false positive onsets. And when a new note of a tuned instrument begins, the output band values will increase significantly because the the main energy of its harmonics will be concentrated in the center frequencies of the semitone bands. This means that the system is specially sensitive to frequency variations that are larger than one semitone.

This way, the spectrum variations produced at the beginning of the notes are emphasized and those produced while the notes are sounding are minimized. This makes the system robust against smooth vibratos that are not higher than a semitone. It also has a special feature; if a pitch bend (*glissando*) occurs, a new onset is usually detected when it reaches more than one quarter tone higher or lower than the starting pitch. This kind of detector can be useful for some music transcription systems, those that have the pitch units measured in semitones.

2 Input Data

2.1 Spectral Analysis

From a digital audio file a short-time Fourier transform (STFT) is computed, providing its spectrogram. In order to remove unused frequency components and increasing spectral resolution downsampling from 44,100 Hz to 22,050 Hz sampling rate was done. Thus, the highest possible frequency is $f_s/2 = 11{,}025$ Hz, which is high enough to cover the range of useful pitches.

The STFT is calculated using a Hanning window with $N = 2048$ samples. An overlapping percentage of 50% ($O = 0.5$) is also applied in order to retain the information at the frame boundaries. The time resolution Δt can be calculated as:

$$\Delta t = \frac{(1 - O)N}{f_s} \quad . \tag{1}$$

Therefore, with the parameter values described, Eq. 1 yields $\Delta t = 46.4$ milliseconds and the STFT provides 1024 frequency values with a spectral resolu-

[1] A "partial" is any of the frequencies in a spectrum, being "harmonic" those multiples of a privileged frequency called fundamental that provides the pitch of the sounding note.

tion of 10.77 Hz. Concert piano frequencies range from $G\sharp_{-1}$ (27.5 Hz) to C_7 (4186 Hz). We want to use 1/12 octave bands. The band centered in pitch $G\sharp_0$ has a center frequency of 51.91 Hz, and the fundamental frequency of the next pitch, A_0, is 55.00 Hz, so a spectral resolution of 10.77 Hz is not enough to build the lower bands.

To minimize this problem, zero padding was applied for having more points in the spectrum, appending three windows of 2048 zero samples at the end of the input signal in the time domain before doing the STFT. Zero padding does not add spectral resolution, but interpolates. With these values, a resolution of $10.77/4 = 2.69$ Hz is obtained.

2.2 Semitone Bands

In this work, the analysis is performed by a computer software in the frequency domain. Therefore, the FFT algorithm is utilized to compute the narrowband (linear) frequency spectrum. Then, this spectrum is apportioned among the octave bands to produce the corresponding octave spectrum, simulating the response of a 1/12 octave filterbank in the frequency domain.

The spectral bins obtained after the STFT computation are analyzed into B bands in a logarithmic scale ranging from 50 Hz (pitch $G\sharp_0$) to 10,600 Hz (pitch F_8), almost eight octaves. This way, $B = 94$ spectral bands are obtained and their center frequencies correspond to the fundamental frequencies of the 94 notes in that range.

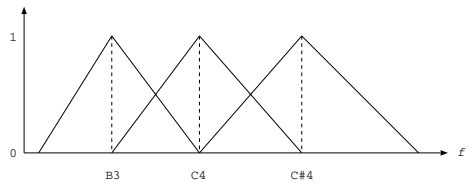

Fig. 1. Example of triangular windowing for pitches B_3, C_4 and $C\sharp_4$

To build the 1/12 octave bands, a set of different sized triangular windows are used (see Fig. 1). There is one window centered at the fundamental frequency of each pitch. For wider windows (those centered in the highest frequencies), many bins are considered but for lower bands only a few bins are used. Therefore, if the input signal is an uniformly distributed noise, wider bands will have higher values than narrower ones. To minimize this problem, a RMS (Root Mean Square) computation is performed, in order to emphasize the highest spectrum values. A simple equation to get each band value $b_k(t)$ at time t can be used;

$$b_k(t) = \sqrt{\sum_{j=1}^{W_k} (X(j,t)w_{kj})^2} \ , \tag{2}$$

being $\{w_{kj}\}_{j=1}^{W_k}$ the triangular window values for each band, W_k the size of the k-th window and X the set of spectrum bins corresponding to that window at time t, with j indexing the frequency bin.

The RMS of the bands is used instead of the energy. This is because small variations in the highest amplitude bands are emphasized, causing false onsets during the sustain stage of some notes. Moreover, some soft onsets could be masked by strong onsets.

3 Note Onset Recognition

3.1 Basic Note Onset Recognition

Like in other onset detection algorithms [2][4][6][7], a first order derivative function is used to pick potential onset candidates. In this work the derivative $c(t)$ is computed for each band k.

$$c_k(t) = \frac{d}{dt} b_k(t) \tag{3}$$

We must combine onset components to yield the onsets in the overall signal. In order to detect only the beginnings of the notes, the positive first order derivatives of all the bands are summed at each time. The negative derivatives are not considered.

$$a(t) = \sum_{k=1}^{B} \max\{0, c_k(t)\}. \tag{4}$$

To normalize the onset detection function, the overall sum of the band values $s(t)$ is also computed:

$$s(t) = \sum_{k=1}^{B} b_k(t) \tag{5}$$

and the sum of the positive derivatives $a(t)$ is divided by the sum of the band amplitudes $s(t)$ to compute a relative difference. Therefore, the onset detection function $o(t) \in [0, 1]$ is:

$$o(t) = \frac{a(t)}{s(t)} \ . \tag{6}$$

The Fig. 2 shows an example of the detection function $o(t)$ for a Mozart real piano melody[2].

A silence threshold μ is applied, in such a way that if $s(t) < \mu$, then $o(t) = 0$. This is done to avoid false positive onsets when the overall amplitude is very low.

The peaks in $o(t)$ are considered as onset candidates and a low level threshold θ is applied to decide which of these candidates are onsets. Due to the fact that

[2] RWC-MDB-C-2001 No. 27 from RWC database [8].

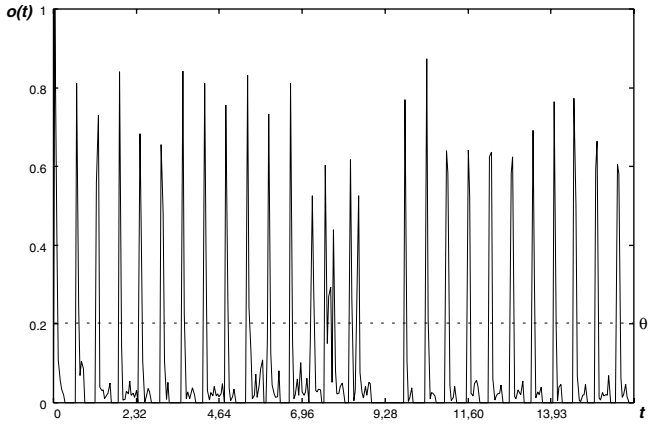

Fig. 2. Example of the onset detection function $o(t)$ for a piano melody. All the detected onsets (peaks over the threshold θ) correspond to actual onsets.

only the peaks are taken into account for onset candidates, two consecutive onsets at t and $t+1$ cannot be detected so the minimum difference in time between two onsets is $2\Delta t = 92.8$ milliseconds.

The human ear cannot distingish between two transients less than 10 ms apart [9]. However, in an onset detector, correct matches usually imply that the target and detected onsets are within a 50 ms window, to allow for the inaccuracy of the hand labelling process [3]. The presented system uses a 46.4 ms window to detect onsets, which is an admisible temporal resolution.

3.2 Note Onset Recognition for Complex Instruments

The previous methodology yields good results for instruments like piano or guitar, having sharp attack envelopes. But for instruments that have a longer attack time, like a church organ, or those with "moving" harmonics as some kind of strings or electric guitars, more time frames should be considered.

The methodology in this case is the same as in the previous subsection, but Eq. 3 is replaced by this one:

$$\tilde{c}_k(t) = \sum_{i=1}^{C} i \cdot [b_k(t+i) - b_k(t-i)] \ , \tag{7}$$

being C the number of considered time frames. This is a variation of an equation (Eq. 5.16) proposed by Young *et al.* in [10] to enhance the performance of a speech recognition system.

The idea of the weighting is that the difference is centered on each particular frame, thus two-side difference (with $C = 1$) is used instead of the frame itself. When using $C = 2$, the difference is calculated from a longer period, playing i the role of a weight.

An example of the onset detection function for a cello melody[3] is shown in Fig. 3 without considering additional frames (a), with $C = 1$ (b) and with $C = 2$ (c).

Note that the higher C is, the lower is the precision in time for detecting onsets but the system yields better results for complex instruments. For a robust detection, the notes need to have a duration $l \geq \Delta t(C + 1)$. If $C = 2$ and with the utilized parameters, $l = 139.2$ ms, so this method variation is not suitable for very rapid onsets[4].

To normalize $o(t)$ into the range $[0, 1]$ Eq. 5 is replaced by

$$\tilde{s}(t) = \sum_{k=1}^{B} \sum_{i=1}^{C} i \cdot b_k(t + i) \tag{8}$$

when the Eq. 7 is used, because only local loudness is considered in Eq. 5.

4 Results

In this work, the experiments were done using an onset detection database proposed by Leveau *et al.* [11] in 2004. Most of its melodies belong to the RWC database [8].

Rigorous evaluation of onset detection is a complex task [12]. The evaluation results of onset detection algorithms presented in various publications are in most cases not comparable [13], and they depend very much on the database used for the experiments. Unfortunately, at the moment there are not similar works using the Leveau *et al.* database, so in this paper our algorithm is not compared with others. However, our system is currently being evaluated at the MIREX 2005 competition[5], which results will be released soon.

A set of real melodies was used to carry out the experiments. To test the system, some real melodies were selected and listened to detect the actual onsets. New audio files were generated adding "click" sounds where the onsets were detected. The number of false positive and negative onsets was finally counted by analysing the generated wavefiles.

The error metric can be defined in precision/recall terms. The precision is the percentage of the detected onsets that are correct. The recall is the percentage of the true onsets that were found with respect to the actual onsets. A false positive is considered as a detected onset that was not present in the signal, and a false negative as an undetected onset.

The silence threshold μ is not very relevant, because in most of the melodies the values of $s(t)$ are usually over this threshold. It is only useful when silences occur or when the considered spectrogram has a very low loudness, so the system is not very sensitive to the variation of this parameter. The threshold θ can control the precision/recall deviation.

[3] RWC-MDB-C-2001 No. 36 from RWC database [8].
[4] 139 ms is the length of a semiquaver when tempo is 107 bpm.
[5] 2nd Annual Music Information Retrieval Evaluation eXchange.

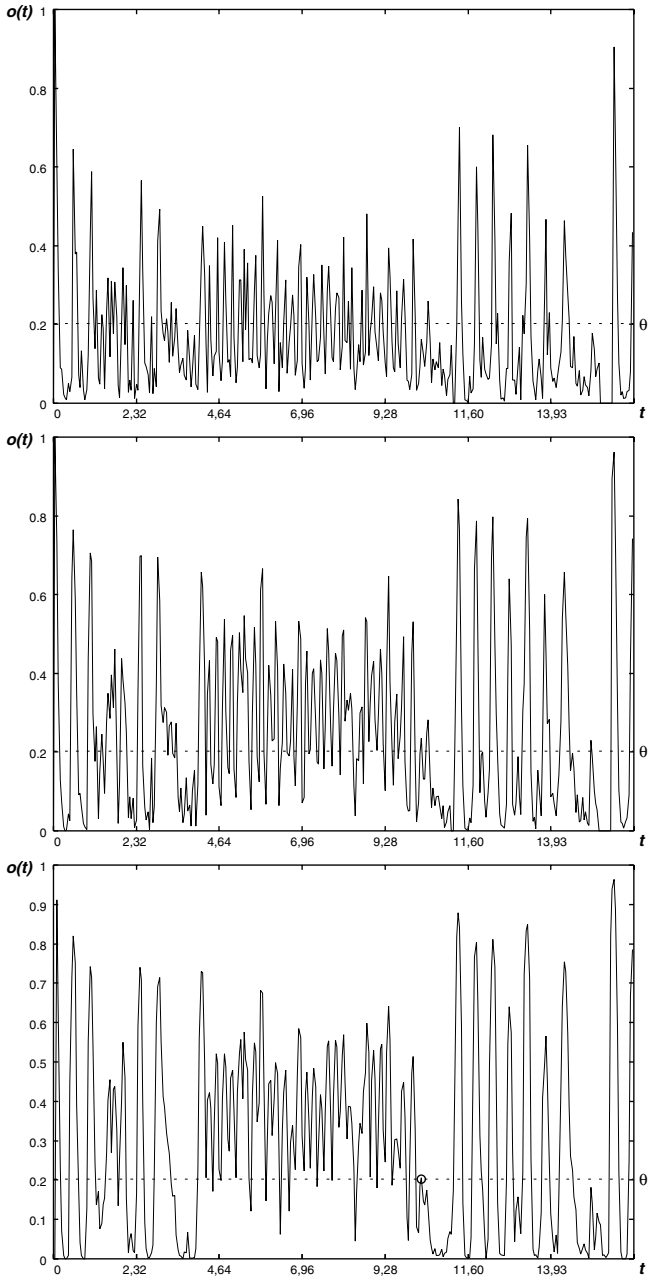

Fig. 3. Onset detection function $o(t)$ for a polyphonic cello melody. (a) Without additional frames; (b) with $C = 1$; (c) with $C = 2$. When $C = 2$, all the onsets were succesfully detected except by one (marked with a circle).

Table 1. Results for the database proposed in [11]. The first columns are the melody name, the duration (secs.), and the number of actual onsets. The next columns are the number of correcly detected onsets (OK), false positives (FP), false negatives (FN), precision (P) and recall (R). The experiments were performed without additional frames (basic detection) and with $C = 2$.

Tested melodies			**Basic detection**					**With C=2**				
Content	Dur (s)	On	OK	FP	FN	P(%)	R(%)	OK	FP	FN	P(%)	R(%)
Solo trumpet	14	60	57	1	3	98.3	95					
Solo clarinet	30	38	38	1	0	97.4	100					
Solo saxophone	12	10	10	4	0	71.4	100					
Solo synthetic bass	7	25	25	1	0	96.2	100					
Solo cello	14	65	49	23	16	68.1	75.4	50	5	15	90.9	76.9
Solo violin	15	79	72	12	7	85.7	91.1					
Solo distorted guitar	6	20	20	3	0	87	100					
Solo steel guitar	14	58	58	2	0	96.7	100					
Solo electric guitar	15	35	31	4	4	88.6	88.6					
Solo piano	15	20	20	0	0	100	100					
Techno	6	56	38	1	19	97.4	67.9					
Rock	15	62	62	21	1	74.7	98.4					
Jazz (octet)	14	52	40	1	12	97.6	76.9					
Jazz (contrabass)	11	52	51	6	1	89.5	98.1					
Classic 1	20	50	49	17	1	74.2	98	50	5	0	90.9	100
Classic 2	14	12	11	15	1	42.3	91.7	11	20	1	35.5	91.7
Pop 1	15	38	32	11	6	74.4	84.2					

4.1 Results Without Additional Frames

The results of the experiments with basic detection are shown in the table 1. They were obtained with a silence threshold $\mu = 70$ and with $\theta = 0.18$.

The system works specially well for the piano melody. In other tested piano melodies results showed that the system is robust for this instrument. It also works well for the tested melodies played by a trumpet, a clarinet, a bass or guitars.

In the melody played by a saxophone a few extra onsets appeared close to the actual onsets. This is due to the nature of this instrument; its attack begins with a small amount of noise, specially evident when it is played legato, like in the tested melody. Its pitch also starts in a frequency slighly lower than the played pitch and it takes a little time to reach the desired pitch. So in some notes both the attack and the moment when the pitch was reached were detected, yielding some false positive onsets.

The cello is a very difficult instrument for onset detection, and the results were not very good when no aditional frames were utilized. Though the violin is another problematic instrument, the results were not bad. Usually, distorted guitars are also a difficult problem for onset detection, but the tested melody yielded good results. More experiments were done with other melodies played by distorted guitars and the system yielded good results too.

In the techno melody, some onsets were not detected probably because they were masked by the strength of the drums. In the rock melody, several false positives appeared when the distorted guitar was played muted. However, in other similar rock melodies the obtained results were very good due to the presence of drums, that are usually helpful for detecting onsets.

The octet jazz melody yielded some false negatives, but most of them belong to very soft onsets produced by the hi-hat of the drumset. The results for the other jazz melody were satisfactory.

In the first classic melody the system obtained good results for the initial notes but, when the singer started, several false positive were achieved. This also happened in another tested singing melodies. The human voice behaviour is different to most of the instruments because of its complex spectral features, so this system do not seem to be the most adequate to deal with this problem.

The second classic melody was very difficult due to the presence of strings, and when no additional frames were considered several false positives appeared. Finally, the pop melody yielded false positives with human voice, and some false negatives corresponding to very soft onsets.

4.2 Results with Additional Frames

As discussed before, for some kind of instruments, like a cello or a church organ, more time frames are needed. In the tested database only three melodies suggest to use additional frames. They are the cello melody and the two classic melodies, and the results with $C = 2$ are in the Tab. 1. The detected onsets considering $C = 1$ were similar to those obtained with basic detection, so they are not shown in the table.

The results with $C = 2$ are not shown for melodies which instrument features do not suggest the use of additional frames. These results are obviously worse considering more time frames than without additional time frames.

The system yielded much better results for the cello and the first classic melodies. However, worse results were obtained for the second classic melody. Obviously, only three examples are not enough to test the performance of the system when $C = 2$ but, unfortunately, in this database only these melodies recommend the use of more frames. In other tested melodies from the RWC database the results improved importantly, for example for the cello melody (in Fig. 3), for an organ and for some classic melodies.

Anyway, in most cases the system yields better results without considering time frames, and more frames should only be utilized for specific instruments.

5 Conclusions and Future Work

In this work, a musically motivated onset detection system is presented. In its basic version, the spectrogram of a melody is performed and 1/12 octave band filters are applied. The derivatives in time are computed for each band and summed. Then, this sum is normalized dividing it by the sum of the band values

in the considered time frame. Finally, all the peaks over a threshold are detected onsets. A simple improvement was made by using more time frames in order to make the system more robust for complex instruments.

The system is intended for tuned musical instruments, and the results for these kind of melodies were very satisfactory. It does not seem to be the most adequate for voice or drums, because it is based in the harmonic properties of the musical instruments. However, when drums were present in the tested melodies, the system was robust. With voice, results are worse due to its harmonic properties.

The simplicity of the system makes it easy to implement, and several future work lines can be developed over this basic scheme. An adaptative filterbank could be added for non-tuned instruments, detecting the highest spectrum peak and moving the fundamental frequency of the closest band to that peak.

A dynamic value of C (the number of aditional time frames) depending on the instruments could also be considered. Usually, in the melodies where C must be increased, the detected onsets in $o(t)$ have lower values than they should have. As an example, in Fig. 2 the peaks detected as onsets have higher values than those detected in Fig. 3 (a). This is because cello attacks are softer than piano attacks. Therefore, the analysis of the $o(t)$ function in the first time frames could be performed to tune the value of C.

Acknowledgements

This work has been funded by the Spanish CICYT project TIRIG with code TIC2003-08496-C04, partially supported by European Union-FEDER funds, and the Generalitat Valenciana project with code GV04B-541. Thanks to Jasón Box for migrating the onset detector C++ code into D2K.

References

1. Klapuri, A. "Sound Onset Detection by Applying Psychoacoustic Knowledge", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, March 15-19, 1999, Phoenix, USA, pp. 3089-3092
2. Goto, M. and Muraoka, Y. "Beat tracking based on multiple-agent architecture — A real-time beat tracking system for audio signals —" in *Proc. of the Second Int. Conf. on Multi-Agent Systems*, pp.103–110, December 1996.
3. Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. and Sandler, M.B. "A tutorial on onset detection in music signals", in *IEEE Transactions on Speech and Audio Processing*", vol. 13, issue 5, pp. 1035 – 1047, Sept. 2005.
4. Scheirer, E.D. "Tempo and beat analysis of acoustic musical signals" *J. Acoust. Soc. Am.*, vol. 103, no.1, pp. 588-601, Jan 1998
5. Duxbury, C., Sandler, M. and Davies, M. "A hybrid approach to musical note onset detection" in *Proc. Digital Audio Effects Conference (DAFX)*, 2002.
6. Goto, M. and Muraoka, Y. "A Real-Time Beat Tracking System for Audio Signals" *Proc. of the 1995 Int. Computer Music Conference*, pp. 171–174, Sep 1995

7. Bilmes, J. "Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning and Reproducing Expressive Timing in Percusive Rhythm". MSc Thesis, MIT, 1993.
8. Goto, M. "RWC music database", published at http://staff.aist.go.jp/m.goto/RWC-MDB/
9. Moore, B.C.J., "An introduction to the Psychology of Hearing", Academic Press, fifth edition, 1997.
10. Young, S., Kershaw, D, Odell, J., Ollason, D., Valtchev, V. and Woodland, P. "The HTK book (for HTK version 3.1)" *Cambridge University*, 2000.
11. Leveau, P., Daudet, L. and Richard, G. "Methodology and tools for the evaluation of automatic onset detection algorithms in music", *Proc. of the Int. Symposium on Music Information Retrieval (ISMIR)*, Barcelona, 2004.
12. Rodet, X., Escribe, J. and Durignon, S. "Improving score to audio alignment: Percussion alignment and Precise Onset Estimation" *Proc. of the 2004 Int. Computer Music Conference*, pp. 450–453, Nov. 2004.
13. Lerch, A., Klich, I. "On the Evaluation of Automatic Onset Tracking Systems", *White Paper, Berlin, Germany, April 2005.*