AN IMPLEMENTATION OF MULTI-BAND ONSET DETECTION

Julien Ricard
julien.ricard@gmail.com

ABSTRACT

This paper describes an algorithm for onset detection based on the method described by Anssi Klapuri in [1]. The signal is first normalized and filtered by a model of the outer ear, and decomposed in frequency bands by gammatone filters. The amplitude envelope is extracted from each band by full-wave rectification and low-pass filtering. Candidate onset positions in each band are estimated by picking the peaks above a threshold on the derivative of the smoothed log-amplitude envelope. A weight corresponding to the maximum value of the smoothed log envelope of the sound event is assigned to each candidate. Close candidates are combined both in time and across frequency bands, and a threshold is applied to get the final onsets.

Keywords: onset detection

1 INTRODUCTION

Detecting the beginning of sound events, typically musical notes, in an audio stream is an important step in higher-level music analysis task, such as tempo estimation or melody extraction. Several onset detection methods have been proposed in the literature, mainly differing from each other in the detection function used (e.g. energy derivative or phase deviation). A tutorial on onset detection is given in [2] and a comparison of algorithms can be found in [3].

In this paper we will describe a psychoacoustically motivated onset detection based on [1] and we will present and discuss the results obtained at the 1st Music Information Retrieval Evaluation eXchange contest (MIREX 2005).

2 DESCRIPTION

The block diagram of our onset detection is shown in figure 1.

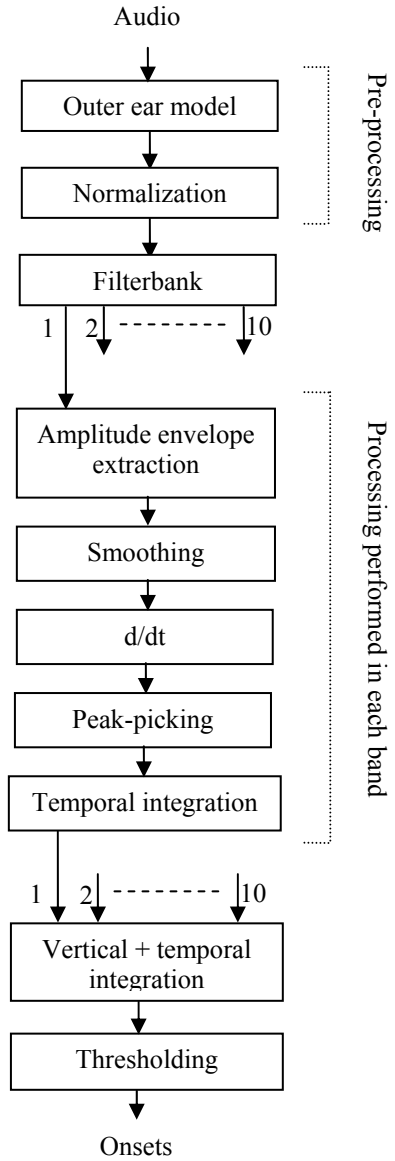

Figure 1 Onset detection block diagram.

2.1 Pre-processing

In the pre-processing step the signal is first filtered by a model of the outer ear described in [4] (eq. 6).

The filtered signal is then normalized according to the value under which lie 90% of the energy values that are not silence (set arbitrarily to values smaller than 8 bits quantization noise). Normalizing the signal allows using a fixed threshold in the peak picking steps.

2.2 Filterbank

The pre-processed signal is then decomposed by a bank of band-pass filters. The filterbank used is a C++ implementation of the ERB cochlear model from Slaney's Auditory Toolbox [5] (we used 10 filters from 100 to 10000 Hz). Decomposing the signal into frequency bands allows detecting onsets that could be masked in full-range complex signal, and detecting

changes in pitch (i.e. note transition) in a constant-amplitude signal.

2.3 Band-wise onset candidate detection

2.3.1 Detection function

The amplitude envelope is extracted in each band by full wave rectification and low-pass filtering, and downsampled to 245 Hz to reduce computation time. The output is then smoothed by a 50 ms half-hanning window, which *preserves sudden changes, but mask rapid modulation* [1]. The derivative of the smoothed log-amplitude envelope is used as the detection function.

2.3.2 Peak picking

Onset candidates are found by picking the highest peak greater than a threshold on each positive segment of the detection function. Each candidate is assigned a weight set to the maximum value of the smoothed log-amplitude envelope of this segment (see figure 2).

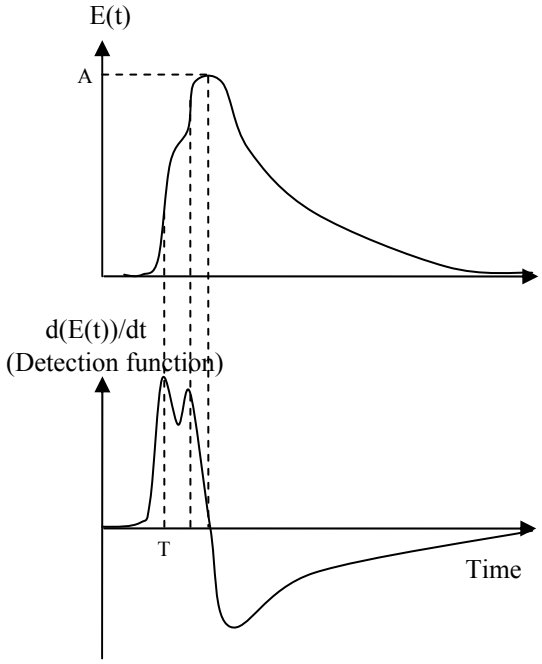

Figure 2 The highest peak in each positive segment of the detection function is selected as an onset candidate. The weight is set to the maximum value of the smoothed log-amplitude envelope of this segment. In this example, the position of the candidate is T and its weight is A.

2.3.3 Temporal integration

In order to model the temporal integration performed by our auditory system (two clicks closer than approximately 50ms cannot be perceived separately), we only keep the candidate having the highest weight in every sequence in which successive candidates are closer than 50 ms (see figure 3).

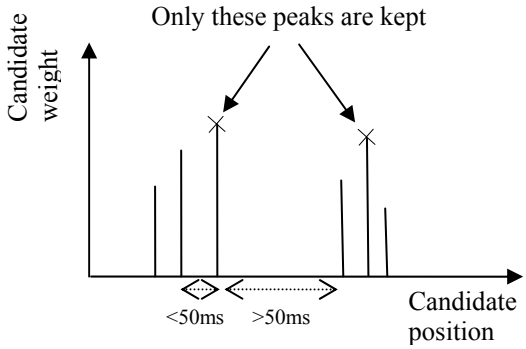

Figure 3 Illustration of the temporal integration process: the candidate having the highest weight in every sequence in which successive candidates are closer than 50 ms is kept.

2.4 Global candidate selection

All the onset candidates detected are combined by summing up their weights across the frequency bands. Temporal integration is performed again by keeping only the candidate with highest weight in every sequence in which successive candidates are closer than 50 ms. Its weight is set to the sum of the weights of all the candidates in the sequence (see figure 4).

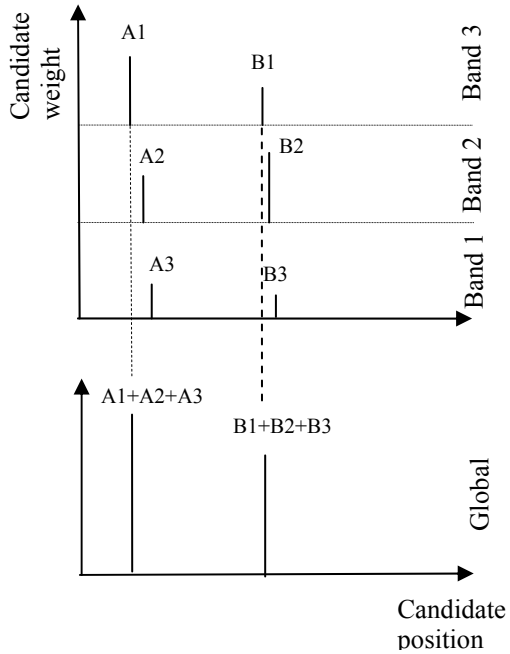

Figure 4 Illustration of the vertical combination and temporal integration performed in order to get the final onset candidates.

A threshold is then applied to the remaining candidates to get the final onsets.

3 EVALUATION AND DISCUSSION

The results of the evaluation performed by the MIREX 2005 team are shown in table 1.

As expected, the algorithm performs poorly for sounds having smooth attacks (e.g. singing voice, sustained strings) and quite well for sounds having sharper attacks (e.g. drums, plucked strings). In order to detect smooth note transitions, the algorithm would probably be improved by including some pitch estimation and a detection of transitions somewhere between stable pitch segments. The small amount, compared to other algorithms, of false positives and doubled onsets motivates the use of a more sensitive detector, by using lower peak picking thresholds for instance. A visual analysis of the errors, i.e. by looking at the detection function, ground-truth onsets and detected onsets, would probably help improve the algorithm and tune the parameters[1].

4 CONCLUSION

The basic idea behind the algorithm described in this paper is to detect energy discontinuities of the signal. Multi-band processing allows detecting these discontinuities even when they are masked in the full-range signal, but smooth note transitions are still not detected. We believe that obtaining in some way the optimal tuning of our algorithm would not improve it significantly. Instead, we believe that an onset detector should not be based only on energy discontinuities and should detect discontinuities of several features (e.g. pitch, phase, timbre…), which implies combining the results of several detection function (see [6] for a segmentation system based on this idea).

ACKNOWLEDGEMENTS

This algorithm was developed with Gilles Peterschmitt at the Music Technology Group (Barcelona, Spain).

REFERENCES

[1] A. Klapuri. “Sound Onset Detection by Applying Psychoacoustic Knowledge", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1999.

[2] J. P. Bello, L. Daudet, S. Abadía, C. Duxbury, M. Davies and M. B. Sandler. “A tutorial on onset detection in music signals”, IEEE Transactions on Speech and Audio Processing. Scheduled for publication on September, 2005.

[3] N. Collins. "A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions". Proceedings of the 118th AES Convention, 2005.

[4] P. Kabal. “Perceptual Evaluation of Audio Quality”, Internal Report, version 2. Department of Electrical & Computer Engineering. McGill University, 2003.

[5] M. Slaney. “Auditory Toolbox”, Technical report #1998-010, Interval Research Corporation

[6] S. Rossignol. “Segmentation et Indexation des Signaux Sonores Musicaux”, PhD Thesis, UNiversité Paris 6, 2000.

[1] The tuning of the parameters was done by trial and errors and optimal tuning of the algorithm could be obtained by an automatic exploration of the parameters space.

	Complex	Poly-pitched	Solo bars and bells	Solo brass	Solo drum	Solo plucked string	Solo singing voice	Solo sustained string	Solo wind	Overall
Rank	4	3	4	1	3	6	6	5	7	3
Average F-measure	71.90%	83.26%	87.17%	72.66%	90.97%	77.85%	27.59%	38.45%	38.57%	74.80%
Average Precision	77.51%	90.12%	81.79%	74.24%	96.46%	88.33%	20.71%	71.33%	38.98%	81.36%
Average Recall	68.20%	80.11%	97.00%	71.51%	87.55%	73.71%	45.22%	32.63%	49.38%	73.70%
Average Recall	2465	677	297	184	2590	330	98	305	153	7099
Total False Positives	650	82	18	17	74	42	391	100	207	1581
Total False Negatives	1094	182	27	29	318	101	131	405	113	2400
Total Merged	97	27	6	0	54	6	1	7	2	200
Total Doubled	3	0	0	0	0	0	0	0	0	3
Average Correct	493.60	225.67	99.00	61.33	863.33	110.00	32.67	101.67	51.00	23.97
Average False Positives	130.00	27.33	6.00	5.67	24.67	14.00	130.33	33.33	69.00	5.18
Average False Negatives	218.80	60.67	9.00	9.67	106.00	33.67	43.67	135.00	37.67	7.70
Average Merged	19.40	9.00	2.00	0.00	18.00	2.00	0.33	2.33	0.67	0.63
Average Doubled	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

Table 1 Results of the MIREX 2005 evaluation for the onset detection described in this paper