SOUND ONSET DETECTION BY APPLYING PSYCHOACOUSTIC KNOWLEDGE

Anssi Klapuri

Signal Processing Laboratory, Tampere University of Technology
P.O.Box 553, FIN-33101 Tampere, FINLAND
`klap@cs.tut.fi`

ABSTRACT

A system was designed, which is able to detect the perceptual onsets of sounds in acoustic signals. The system is general in regard to the sounds involved and was found to be robust for different kinds of signals. This was achieved without assuming regularities in the positions of the onsets. In this paper, a method is first proposed that can determine the beginnings of sounds that exhibit onset imperfections, i.e., the amplitude envelope of which does not rise monotonically. Then the mentioned system is described, which utilizes band-wise processing and a psychoacoustic model of intensity coding to combine the results from the separate frequency bands. The performance of the system was validated by applying it to the detection of onsets in musical signals that ranged from rock to classical and big band recordings.

1. INTRODUCTION

Onset detection plays an important role in the computational segmentation and analysis of acoustic signals. It greatly facilitates cut-and-paste operations and editing of audio recordings. The onset information may also be used in audio/video synchronization and timing, or passed for further analysis and recognition for example in an acoustic supervision system.

We use the term *onset detection* to refer to the detection of the beginnings of discrete events in acoustic signals. A percept of an onset is caused by a noticeable change in the intensity, pitch or timbre of the sound [1]. A fundamental problem in the design of an onset detection system is distinguishing genuine onsets from gradual changes and modulations that take place during the ringing of a sound. This is also the reason why robust one-by-one detection of onsets has proved to be very hard to attain without significantly limiting the set of application signals.

A lot of research related to onset detection has been carried out in recent years. However, only few systems have set out to solve the problem of one-by-one onset detection [1][2][3]. Instead, most systems aim at higher-level information, such as the perceived *beat* of a musical signal [4][5][6], in which case long-term autocorrelations and regularities can be used to remove single errors and to tune the sensitivity of the low-level detection process.

In this paper, we first propose a mathematical method to cope with sounds that exhibit onset imperfections, i.e., the amplitude envelope of which rises through a complex track and easily produces erroneous extra onsets or an incorrect time value. Then we propose the application of psychoacoustic models of intensity coding, which enable us to determine system parameters that apply to a wide variety of input signals. This allows processing them without a priori knowledge of signal contents or separate tuning of parameters.

The realized system was validated by applying it to the detection of onsets in musical signals. This was done mainly for two reasons. First, musical signals introduce a rich variety of sounds with a wide range of pitches, timbres and loudnesses. Different combinations of onsetting and backgrounding sounds are readily available. Second, verifying the contents of a musical signal is somewhat easier than in the case of environmental sounds. Also the concept of a perceivable onset is better defined. It should be noted, however, that the algorithm is not limited to musical signals, because the regularities and rhythmic properties of musical signals are not utilized in the detection process. The system performs reliably for input signals that ranged from rock music to classical and big band recordings, both with and without drums.

2. SYSTEM OVERVIEW

The earliest onset detection systems typically tried to process the amplitude envelope of a signal as a whole (see e.g. [7]). Since this was not very effective, later proposals have evolved towards band-wise processing. Scheirer was the first to clearly point out the fact that an onset detection algorithm should follow the human auditory system by treating frequency bands separately and then combining results in the end [4]. An earlier system of Bilmes's was on the way to the same direction, but his system only used a high-frequency and a low-frequency band, which was not that effective [2].

Scheirer describes a psychoacoustic demonstration on beat perception, which shows that certain kinds of signal simplifications can be performed without affecting the perceived rhythmic content of a musical signal [4]. When the signal is divided into at least four frequency bands and the corresponding bands of a noise signal are controlled by the amplitude envelopes of the musical signal, the noise signal will have a rhythmic percept which is significantly the same as that of the original signal. On the other hand, this does not hold if only one band is used, in which case the original signal is no more recognizable from its simplified form.

The overview of our onset detection system is presented in Figure 1. It utilizes the band-wise processing principle as motivated above. First, the overall loudness of the signal is normalized to 70 dB level using the model of loudness as proposed by Moore et al. [8]. Then a filterbank divides the signal into 21 non-overlapping bands. At each band, we detect *onset components* and determine their time and intensity. In final phase, the onset components
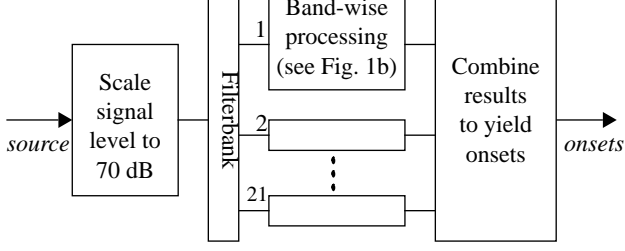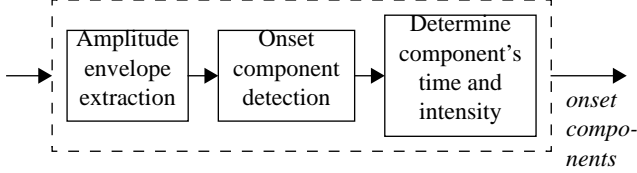
Figure 1a. System overview.

Figure 1b. Processing at each frequency band.

Dashed line:
$$\frac{d}{dt}(A(t))$$
Solid line:
$$\frac{d}{dt}(\log(A(t)))$$
where $A(t)$ denotes the amplitude envelope function.

Figure 2. Onset of a piano sound. First order *absolute* (dashed) and *relative* (solid) difference functions of the amplitude envelopes of six different frequency bands.

are combined to yield onsets.

Since we use psychoacoustic models both in onset component detection, in its time and intensity determination, and in combining the results, it is important to use a filterbank which can provide input to the models. Therefore, we choose a bank of nearly critical-band filters which covers the frequencies from 44 Hz to 18 kHz. The lowest three among the required 21 filters are one-octave band-pass filters. The remaining eighteen are third-octave band-pass filters. All subsequent calculations can be done one band at a time. This reduces the memory requirements of the algorithm in the case of long input signals, assumed that parallel processing is not desired.

The output of each filter is full-wave rectified and then decimated by factor 180 to ease the following computations. Amplitude envelopes are calculated by convolving the band-limited signals with a 100 ms half-Hanning (raised cosine) window. This window performs much the same energy integration as the human auditory system, preserving sudden changes, but masking rapid modulation [9][4].

3. CALCULATION OF ONSET COMPONENTS

3.1 Onset Component Detection

Several algorithms for picking potential onset candidates from an amplitude envelope function have been presented in the literature [5][6][2][4]. Despite the number of variants, practically all of them are based on the calculation of a first order difference function of the signal amplitude envelopes and taking the maximum rising slope as an onset or an onset component.

In our simulations, it turned out that the first order difference function reflects well the loudness of an onsetting sound, but its maximum values fail to precisely mark the time of an onset. This is due to two reasons. First, especially low sounds may take some time to come to the point where their amplitude is maximally rising, and thus that point is crucially late from the physical onset of a sound and leads to an incorrect cross-band association with the higher frequencies. Second, the onset track of a sound is most often not monotonically increasing, and thus we would have sev-
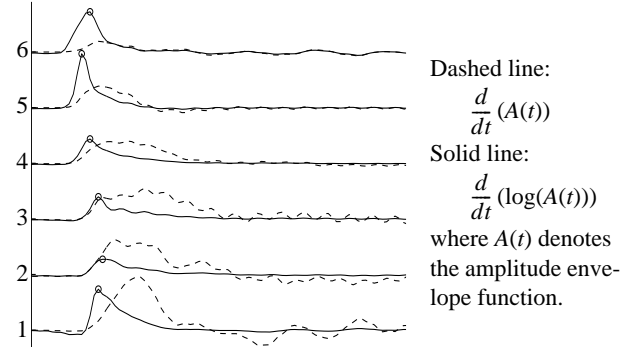
eral local maxima in the first order difference function near the physical onset (see plots with a dashed line in Figure 2).

We took an approach that effectively handles both of these problems. We begin by calculating a first order difference function

$$D(t) = \frac{d}{dt}(A(t)),$$

where $A(t)$ denotes the amplitude envelope function. $D(t)$ is set to zero where signal is below minimum audible field. Then we divide the first order difference function by the amplitude envelope function to get a first order *relative difference function W*, i.e., the amount of change in relation to the signal level. This is the same as differentiating the logarithm of the amplitude envelope.

$$W(t) = \frac{d}{dt}(\log(A(t)))$$

We use the relative difference function $W(t)$ both to detect onset components and to determine their time. This is psychoacoustically relevant, since perceived increase in signal amplitude is in relation to its level, the same amount of increase being more prominent in a quiet signal. According to Moore, the smallest detectable change in intensity is approximately proportional to the intensity of the signal [10]. That is, $\Delta I / I$, the Weber fraction, is a constant. This relationship holds for intensities from about 20 dB to about 100 dB above the absolute threshold. The function $\Delta I(t) / I(t)$ is equivalent to $W(t)$, since the frequency f in $I(t) = A(t) \cdot f$ is reduced in the division. Thus we detect onset components by a simple peak picking operation, which looks for peaks above a global threshold T_{det} in the relative difference function $W(t)$.

The relative difference function effectively solves the abovementioned problems by detecting the onset times of low sounds earlier and, more importantly, by handling complicated onset tracks, since oscillations in the onset track of a sound do not matter in relative terms after its amplitude has started rising. To clarify this, we plotted the absolute and relative difference functions of the onset of a piano sound in Figure 2. Both of the benefits discussed can be seen clearly.

3.2 Intensity of an Onset Component

Simultaneously occurring sounds combine by a linear summation. In determining the intensity of an already detected onset component, we can assume the level of backgrounding sounds to be

momentarily steady and take the increase in level to be due to the onsetting sound(s). Thus the asked intensity can be picked from the first order difference function $D(t)$, multiplied by the band center frequency f_B. The intensity is needed later when onset components are combined to yield onsets of the overall signal.

An appropriate point in time to pick the intensity from $D(t)$ is not as early as where the onset was determined to occur. Instead, we scan forward up to the point where amplitude envelope starts decreasing and determine the intensity at the point of maximum slope, i.e., at the maximum value of $D(t)$ between the onset and the point where amplitude stops increasing.

After intensities has been determined for all onset components at the band, we check them through and drop out components that are closer than 50 ms to a more intense component. Remaining ones are accepted.

4. COMBINING THE RESULTS FROM THE BANDS

In the final phase we combine onset components from separate bands to yield onsets of the overall signal. For this purpose, we implemented the model of loudness as proposed by Moore, Glasberg and Baer [8]. Input to our implementation is a vector of sound intensities at third-octave bands between 44 Hz and 18 kHz, from which the program calculates the loudness of the signal in phons. To optimize the computational efficiency of the procedure, we slightly simplified the model by making the shape of the excitation pattern, i.e., the intensity spread between adjacent critical bands independent from sound pressure level. This accelerated the computations remarkably, but did not make a significant difference to the estimated loudness values for the sound intensity levels we are using.

The onsets of the overall signal are calculated as follows. First the onset components from different bands are all sorted in time order, and are regarded as sound onset candidates hereafter. Then each onset candidate is assigned a loudness value, which is calculated by collecting onset components in a 50 ms time window around the candidate and feeding their intensities to the corresponding frequency bands of the loudness model of Moore et al. Since most candidates have only a couple of contributing onset components at different bands, we must use minimum level, or background noise level for the other bands in the input of the model. Repeating this procedure to each onset candidate yields a vector of candidate loudnesses as a function of their times, as illustrated in Figure 3 for a popular music signal.

Onset loudnesses that were estimated using the abovementined procedure corresponded very well to the perceived loudnesses of the onsets in verificative listening tests. It turned out that a robust detection of onsets in very diverse kinds of signals can now be achieved by a simple peak picking operation, which looks for onset candidates above a global threshold value T_{final}. We drop out onset candidates whose loudness falls below the threshold. Then we also drop out candidates that are too close (50 ms) to a louder candidate. Among equally loud but too close candidates, the middle one (median) is chosen and the others are abandoned. The remaining onset candidates are accepted as true ones. A good value for T_{final} was found to be 25 dB for signals, whose average loudnesses had been normalized to 70 dB level.
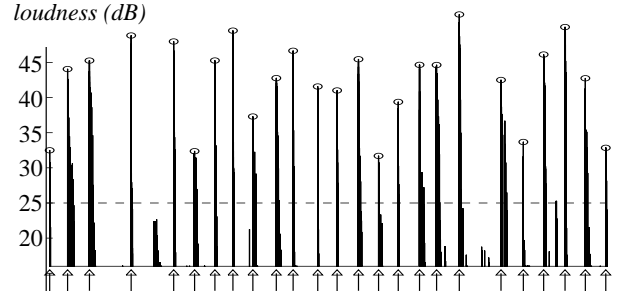

loudness (dB)

Figure 3. The loudness of onsets as a function of their time. The genuine onsets can now be quite easily discerned.

5. VALIDATION EXPERIMENTS

The presented procedure was verified by testing its performance in detecting onsets in musical signals. The signals were selected to comprise a large variation of musical instruments and a wide dynamic and pitch range. Signals both with and without drums were included. Another goal was to include representative excerpts from different musical genres, ranging from jazz and rock to classical and big band music.

Approximately ten second excerpts were sampled from each performance. These periods were carefully inspected and their onset times were marked. The excerpts were then feeded to the onset detection system and its results were compared to the manual transcription. All simulation cases were computed using the very same set of parameter values and thresholds, without separate tailoring for each simulation case. The algorithm itself was as explained above. Higher-level rhythmic properties and regularities of musical signals were not utilized in the detection.

It is interesting to note that the limitations of our detection system resemble those of human perception. We define a *pseudo-onset* to be a sound beginning, which undisputably exists in a signal, but cannot be detected by a human listener if the signal is not presented in short segments and several times. Since objective listening test could not be arranged, we regard undetected pseudo-onset as errors, too. It turned out that the detection of some pseudo-onsets could not be achieved without giving rise to several erroneous extra onsets that are due to gradual changes and modulations during the ringing of sounds.

Onset detection results for ten different musical signals are summarized in Table 1. The total number of onsets, number of undetected onsets and the number of erroneous extra onsets are given. A measure of correctness in the rightmost column is calculated as

$$correct = \frac{total - undetected - extra}{total} \cdot 100 \,\%.$$

A more detailed discussion of each case follows.

Chopin's classical piano etude (op. 25, no. 4) was a trivial case. Still three onsets fell below threshold because the notes were low pitched, played softly and masked by other notes. *Al Di Meola*'s 'Orient Blue' represents a much more difficult case. The piece is polyphonic and employs the whole dynamic and pitch range of the acoustic guitar. Shortest inter-note intervals are only a fifteenth of a second. Good results were achieved partly because of

Table 1: Summary of onset detection results.

signal	worth notice in contents	onsets in total	unde-tected	extra	correct (%)
Chopin	acoustic piano	59	3	–	95
AldiMeola	acoustic guitar	62	5	1	92
Police	singing, el.guitar, drums	49	4	1	90
U2	el. guitar rif, distorted	19	1	2	84
Grusin	piano, percussion, drums	51	3	–	94
MDavis	brasses, double-bass	34	2	1	91
Miller	big band	46	5	1	87
Bach	chamber ensemble	51	3	1	92
Vivaldi	symphony orchestra	33	7	10	48
Beethoven	symphony orchestra	30	–	28	7

the absense of noise and other instruments.

Police's 'It's Alright for You' is from rock music genre, dominated in loudness by singing, electric guitars and drums. Onset detection is a success and resembles the results that were derived with other rock-pieces. At some moments singing produced double-onsets for phonem combinations like "-ps-", where both *p* and *s* produce an onset. All of these occurred inside the 50 ms time window, however, and were therefore fused. *U2* is an electric guitar rif, taken from the band's performance of 'Last Night on Earth'. The excerpt is played with distorted sound, without accompanying instruments. This case illustrates that even ambiguos situations, i.e., rough sounds, can be handled. *Grusin*'s 'Punta del Soul' is classified to fusion jazz, but the selected excerpt resembles mostly popular music. Various percussions included were detected without trouble.

Miles Davis's 'So What' introduces a selection of jazz band instruments: a trumpet, tenor and alto saxophones, piano, plucked double-bass and gentle drums. Both brass instrument onsets and soft pluckings of the double bass were consistently detected. *Glen Miller*'s 'In the Mood' is dominated by big band's brass instruments of the performing orchestra. All undetections occurred in a clarinet melody, which was partly masked by louder instruments.

Bach's Brandenburg Concerto was sampled from the performance of Munich Chamber Ensemble, which comprises strings, woodwinds and brass instruments. It is worth notice that onsets were detected even at moments where strings alone were carrying the rhythm and played tying consecutive notes to each other.

As a sharp contrast to the robust detections, all symphony orchestra performances turned out to be resolved very poorly. *Vivaldi*'s 'The Four Seasons' and *Beethoven*'s Symphony No. 5 are given as examples in Table 1. The clear discrepancy with human perception is not due to the type of instruments involved, since they were detected well in smaller ensembles. Instead, two causes are supposed. Firstly, individual physical sound sources can no more be followed in a symphony orchestra, but resulting onsets derive from several sources and are smoothed. Secondly, it was revealed by a certain hammond organ solo that a strong amplitude modulation at the middle frequencies confuses the system. It seems that the human auditory system has a special ability to ignore even a very loud amplitude modulation if it is inconsistent, and to con-

centrate on frequencies where structure is found.

6. CONCLUSIONS

We first discussed problems that arise in the one-by-one detection of sound onsets. Then a system was described, which builds upon the use of relative difference function and application of the psychoacoustic models of intensity coding. This was done in the framework of the band-wise processing idea. Experimental results show that the presented system exhibits a significant generality in regard to the sounds and signal types involved. This was achieved without higher-level logic or a grouping of the onsets. The system introduces only two thresholds that need to be experimentally found, i.e., that are not deduced from psychoacoustic metrics. These thresholds are common to all input signals.

One of the shortcomings of our method lies in its inability to deal with a strong amplitude modulation which is met in classical ensembles and in certain instrumental sounds. In general, the proposed system was well able to discern between genuine onsets and gradual changes and modulations in the sounds themselves. In the case of musical signals, an additional higher-level analysis would still significantly improve the accuracy of the system.

7. REFERENCES

[1] Moelants D., Rampazzo C. "A Computer System for the Automatic Detection of Perceptual Onsets in a Musical Signal". In Camurri, Antonio (Ed.). "*KANSEI, The Technology of Emotion*", pp. 140–146. Genova, 1997.

[2] Bilmes J. "Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm". MSc thesis, Massachusetts Instute of Technology, 1993.

[3] Schloss A. "On the Automatic Transcription of Percussive Music — From Acoustic Signal to High-Level Analysis". Ph.D. thesis, Stanford University, 1985. Report STAN-M-27.

[4] Scheirer E. "Tempo and Beat Analysis of Acoustic Musical Signals". Machine Listening Group, MIT Media Laboratory, 1996.

[5] Goto M., Muraoka Y. "Beat Tracking based on Multiple-agent Architecture - A Real-time Beat Tracking System for Audio Signals". *Proceedings of The Second International Conference on Multiagent Systems*, pp.103–110, 1996.

[6] Goto M., Muraoka Y. "A Real-time Beat Tracking System for Audio Signals". *Proceedings of the 1995 International Computer Music Conference*, pp.171–174, September 1995.

[7] Chafe C., Jaffe D., Kashima K., Mont-Reunaud B., Smith J. "Source Separation and Note Identification in Polyphonic Music". Stanford University, Department of Music, Report STAN-M-29. 1985

[8] Moore B., Glasberg B., Baer T. "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness". *J. Audio Eng. Soc.*, Vol. 45, No. 4, pp. 224–240. April 1997

[9] Todd, McAulay. "The Auditory Primal Sketch: a Multiscale Model of Rhythmic Grouping". *Journal of New Music Research,* 23, pp. 25–70, 1992.

[10] Moore B. (ed). "Hearing". Handbook of Perception and Cognition, 2nd Edition. Academic Press, 1995.