ON THE EVALUATION OF AUTOMATIC ONSET TRACKING SYSTEMS

Alexander Lerch
zplane.development
Berlin, Germany
lerch@zplane.de

Ingmar Klich
Dept. of Communication Research
Technical University of Berlin, Germany
ingmale@cs.tu-berlin.de

ABSTRACT

This paper summarizes the problems, definitions and requirements that are important for the evaluation of onset tracking systems for audio signals in PCM format. Different procedures and metrics for evaluation and parametrization are presented and commented. Overall, a complete methodology for the evaluation of automatic onset detection systems is proposed.

Keywords: evaluation, onset-tracking, methodology

1 INTRODUCTION

The growing number of automatic onset tracking systems in research as well as in commercial applications shows the importance of such systems for the field of music information retrieval (MIR). Extracted onset information for example can be used in applications for the detection of tempo, beat locations, time signature, automatic transcription, automatic alignment as well as the segmentation of audio signals.

Unfortunately the evaluation results of automatic onset detection algorithms presented in various publications are in most cases not comparable. According to Downie [3], the differences in evaluation methods can be ascribed to the lack of familiarity among members of the various domains with traditional IR evaluation techniques, the lack of standardized, multirepresentational test collections and the lack of a standardized set of relevance judgements.

While the existence of standardized test collections is at least partly limited by issues that cannot be easily overcome, the intent of this paper is to summarize previous efforts about onset detection evaluation and to propose a complete evaluation methodology for audio onset detection systems to make the evaluation and results of such systems more systematic and comparable.

Despite the fact that much research is being done in the field of onset detection, the evaluation part in most publications is usually far less elaborate than the algorithmic description itself, and the problems of evaluating such systems are addressed only partly. In many cases, only the number of correct detections is reported, sometimes with a short note of what the definition of a correct detection is. The lack of information about the test procedure and the test signals used makes it nearly impossible to estimate the algorithm's detection performance and to compare the results with other onset detection publications. Only recently the problem of proper evaluation got more in the researcher's focus. Leveau et al. [7] pointed out the difficulties of manual onset annotation for real-world audio signals. In the context of MIREX [9], the Music Information Retrieval Evaluation Exchange, some effort has been made in proposing a standardized test environment for audio onset detection systems.

The specific problems for the evaluation of audio onset tracking systems can be summarized as:

- lack of proper definition of the term onset, i.e. it is not completely clear what is detected by the system and what is the required measurement accuracy
- lack of an adequate amount of test material due to the effort and error-proneness of manual reference onset annotation for the test files
- lack of standardized and critical test material, which is required to make different results comparable
- lack of general evaluation metrics for the presentation of meaningful results
- lack of a generalized test procedure

The first sections of this paper deal with the definition of onset and onset time and the human perception of onsets. Thereafter, the requirements for the evaluation and test signals are proposed.

2 DEFINITION OF ONSET

Usually, onsets are defined as the start of a (musical) sound event, such as the beginning of a tone or the stroke on a percussive instrument. The term onset is frequently used as a synonym to onset time, but it should be more correct to state that its time position (i.e. the onset time) is one (most likely the main) property of the onset, while an onset can have other properties, e.g. its strength.

2.1 About Onset Time

In most cases, the start of a musical sound is not an exact point in time, but a time span called rise time or initial transient time. This is basically the time from the first instrument-induced measurable oscillation until either the quasi-periodic state or a maximum amplitude is reached, although there are other definitions. The rise time can vary significantly between different musical instruments resp. groups of instruments, e.g. from about $5ms$ for some percussive instruments up to $200ms$ for woodwind instruments like flute under certain circumstances [14].

Since the rise time is a time span and the onset time is a point in time, what discrete point in time is then the required onset time? Three different definitions of onset times can generally be distinguished as pointed out by Repp [13]:

1. *Note Onset Time* (*NOT*): the time when the instrument is triggered to make a sound. In the MIDI domain, the *NOT* is exactly the time of the Note-On command. Note that depending on the instrument or sound used, this is not necessarily the time when the signal becomes audible or detectable.
2. *Acoustic Onset Time* (*AOT*): the first time when a signal or an acoustic event is measurable. Sometimes the *AOT* is called *Physical Onset Time*.
3. *Perceptual Onset Time* (*POT*): the first time when the event can be perceived by the listener. The *POT* might also be distinguished from the *Perceptual Attack Time* (*PAT*), the time that is relevant for the rhythmic perception of the sound [5]. While the *PAT* might occur later than the *POT*, they will be equal in many cases. For the sake of simplicity, there will be no distinction between *POT* and *PAT* in the following.

The *POT* never can occur before the *AOT*, which never occurs before the *NOT*. Due to the "perceptual" definition of the *POT*, the exact location cannot be determined easily but has to be measured in a listening test. Gordon [5] and Zwicker [15] found strong location drifts of the *PAT* resp. *POT* depending on the waveform properties during the rise time.

Given the three definitions above, the question arises which of the three onset times should be assumed to be the reference onset time for the evaluation of the onset detection. Due to the symbolic nature of the *NOT*, it simply cannot be measured from the audio signal. The choice between *AOT* and *POT* might be application dependent; assuming that musicians adapt their timing to their sound perception and that most MIR-Systems are trying to analyze the perceptual audio content, the *POT* is most likely the time that is wanted.

2.2 Time Resolution of Onset Perception

In order to estimate the required time accuracy of an onset detection system, the human ability to exactly locate onset times and to distinguish succeeding onsets is of great interest, since most algorithms are targeting to be at least as accurate as the human perception.

Hirsh [6] found that temporal discrimination of two succeeding onsets is possible if the onset time difference is as little as $2ms$. However, in order to determine the order of the stimuli, their distance had to be about $20ms$. The measurements were done with synthetic signals with short rise times.

Gordon [5] reports a standard deviation of $12ms$ for the accuracy of onset times specified by test listeners, using 16 real-world monophonic sounds of different instruments played in an indefinitely long loop pattern with Inter-Onset-Intervals (IOIs) of $600ms$. Friberg and Sundberg [4] undertook a similar experiment using tone stimuli. For IOIs smaller than $240ms$, they reported a just noticeable difference of about $10ms$, and increasing values for larger IOIs.

Repp [12] reported for the manual annotation of onset times by one listener in the context of piano recordings a mean absolute measurement error of about $4.3ms$ and a maximum error of about $35ms$. In a recent investigation, Leveau et al. [7] had three test subjects annotating the onset times in audio files of various genres and instrumentations. The results showed a mean absolute measurement error over all test data of about $10ms$; for one piece of classical music, the mean absolute measurement error nearly reached $30ms$.

Rasch [11] evaluated the onset time differences between instruments in three ensemble performances He found synchronization deviations in a range between $30ms$ and $50ms$ between the (string and woodwind) instruments, while the mean onset time differences were in the range of $\pm 6ms$. However, it is complicated to distinguish between the accuracy of measurement and performance in this case.

It may be concluded that the measurement accuracy highly depends on the used input data. The presented publications imply that a reasonable demand for the detection accuracy of an automatic onset detection system cannot be smaller than in a range of $5 - 10ms$ and has to be as high as $50ms$ for certain signals with music including many instruments and/or instruments with long rise times.

3 EVALUATION PROCEDURE

When evaluating onset detection systems, the following parameters could be taken into account:

- detection performance
- detection accuracy
- robustness for noisy and band limited input signals
- workload of the algorithm

For each of these parameters, the definition of meaningful rating metrics with a predefined range, preferably between 0 and 1, is desirable. The type and amount of test signals has to be specified to make results as comparable as possible.

3.1 Detection Performance

The detection performance is probably the most important value for the evaluation of onset tracking. The extracted onset times have to be compared with previously defined reference onset times. Two possible errors can occur: no onset is detected in the case of a reference onset (false negative: *FN*), and an onset is detected where no refer-

ence onset is available (false positive: *FP*). Both of these measurements assume the definition of a correct detection; usually, a correct detection is assumed to be within a time window of $50ms$ around the reference onset time.

Before the evaluation itself is carried out, the parameters of the onset detection should be adjusted for the desired "working point". The relation O_{FN}/O_{FP} should be near the value 1 if missed and additional onsets are considered to be equally bad. The so-called Receiver Operating Curve (ROC) plots the number of the correct detections with respect to the *FP*s, allowing an intuitive way of adjusting the desired reliability.

Several measurements of detection performance have been proposed in the past. Cemgil et al. [1] proposed the relation of the total number of detections O_t, the number of *FN*s O_{FN} and the total number of reference onsets O_r as a measure of detection performance:

$$q_{cemgil,1} = \frac{O_t - O_{FN}}{O_r} \tag{1}$$

While this is a simple definition of the detection rate, it does not take into account the falsely detected additional onsets, and thus can result in misleading values in the case of many *FP*s.

Liu et al. [8] proposed a similar value for the detection rate, additionally taking into account the number of false detections O_{FP}:

$$q_{liu} = \frac{max(O_t, O_r) - (O_{FN} + O_{FP})}{max(O_t, O_r)} \tag{2}$$

At least theoretically, the result can be negative, which is not desirable for a detection rate measure that should be in the range between 0 and 1.

For that reason, the proposed reliability measurement is a simple measurement of relative error:

$$q = \frac{O_t - (O_{FN} + O_{FP})}{O_r + (O_{FN} + O_{FP})} \tag{3}$$

The resulting value has the desired range between 0 and 1. The number of missing detections has the same weight as the number of false positives. In some contexts it might be desirable to weight O_{FN} and O_{FP} by different values, because one of both is more important than the other. In these cases, a scaling factor λ between 0 and 1 can be introduced that weights the sum of missing and falsely detected onsets: $\lambda \cdot O_{FN} + (1 - \lambda) \cdot O_{FP}$. Then, however, there is the possibility of negative output values as well.

A reliability measurement could also include the time distance between reference and detected onset time. This way, a detected onset would not only be weighted as correct or incorrect but would also be weighted with respect to its correctness. An intuitive way to do so could be to weight the distance $d_{r,t}$ between reference and detected time with a window function $W(d)$. Cemgil et al. [2] proposed such a measure in the context of evaluation of beat-tracking systems with a Gaussian window function $W(d) = exp(-d^2/2\sigma^2)$. Adapted to the onset tracking evaluation problem it would look like:

$$q_{cemgil,2} = \frac{\sum_{\forall r} \max_{\forall t} W(d_{r,t})}{(O_r + O_t)/2} \tag{4}$$

This measurement has again the limitation that it is not able to correctly handle additional onsets.

3.2 Detection Accuracy

While the previous measurement evaluated the number of correct and false detections within a relatively large tolerance window, the goal of this test is to give a detailed overview of the timing accuracy of the evaluated algorithm. The time difference $d_{r,t}$ between reference and detected onset times is measured over all test files. The distribution of the resulting time differences contains all necessary data for timing evaluation; interesting values are the mean value

$$d_{mean} = \sum_{\forall r} d_{r,t}, \tag{5}$$

the standard deviation or a confidence interval

$$\sigma_d = \sqrt{\frac{1}{O_r} \sum_{\forall r} (d_{r,t} - d_{mean})^2}, \tag{6}$$

and the absolute maximum value of the deviation $d_{max} = \max_{\forall r} d_{r,t}$.

Furthermore, a measure of statistical significance like the p-value should be given to attest the reliability of the results.

3.3 Robustness and Workload

The evaluation of robustness and workload of the algorithm may be useful dependent on the target application and are easy to carry out. The robustness against noise and bandwidth limitations can be undertaken straight forward using the test described in section 3.1, but with added noise and/or low pass filtered test signals. Properties of noise and filter depend on the target application; proposed properties are white Gaussian noise of $-20dB$ RMS power and a low pass filter with a cut-off frequency at $10kHz$.

The evaluation of the workload produced by the proposed onset detection algorithm may be of interest to estimate the complexity and real-time capabilities of the system. Since performance measurements can vary even between similar computer configurations, it may only give a rough figure of the algorithms processing performance. Workload measurements usually give the relation between the required computation time t_r and the overall length of the tested audio data of the test data base t_l by calculating t_r/t_l with respect to the used processor.

4 TEST SIGNAL DATABASES

4.0.1 *Detection Performance*

The test signals to evaluate detection performance should be preferably "real world" signals such as signals from CD with onset times annotated per hand. However, as several publications (e.g. [12] and [7]) point out, the manual annotation is a very time-consuming task. Therefore, two alternatives for the generation of test sequences may be considered; natural recordings with a symbolic trigger like recordings of the Yamaha Disklavier and audio data synthesized from symbolic data. In both cases, the symbolic data is available e.g. in the MIDI-format, allowing

the easy automated extraction of *NOT*s. Given the relatively broad range of the tolerance interval of $50ms$, the difference between *NOT* and *POT* can be neglected.

The test database should include the following signals to make the evaluation as general as possible:

- various genres (pop, rock, symphonic, chamber music, electronic, etc.)
- various instrumentations
- different tempi and complexity
- signals including noisy parts and various kinds of tremolo and vibrato since many onset detection algorithms are sensitive to these signal properties resp. performance styles

As mentioned above, audio files with manually annotated onset times are way more difficult to find. This is on the one hand due to intellectual property issues, on the other hand due to the time-consuming task of annotation. To our best knowledge, the only publicly available database for onset tracking evaluation with manually annotated onset times is published online by Leveau et al. [10] and contains several audio files of different genres.

4.0.2 Detection Accuracy

To evaluate detection accuracy, it is desirable to have input signals with a high correct detection rate. Furthermore it should be avoided to use manually annotated reference onset times because the annotation errors may influence the result. Thus, test signals synthesized from MIDI signals should be used. To ensure a minimum influence of the difference between *NOT* and *POT* and to ensure a high detection rate, the usage of electronic signals with an easy detectable rise/attack time of minimal length is suggested.

Files in MIDI-format are easily available online, partly for free. The generation of test signals for the evaluation of timing accuracy should therefore possible without too much problems.

5 CONCLUSIONS

We summarized the general problems in evaluating onset tracking systems and defined the requirements for the evaluation with respect to the required accuracy and the test data. Evaluation metrics as well as a general test procedure were proposed, with the goal to animate researchers to publish more comparable evaluation results. One of the main issues that remains to be accomplished is the establishment of a large test database with manually annotated onset times that is available to the research community.

References

[1] Ali T. Cemgil, Peter Desain, and Bert Kappen. Rhythm Quantization for Transcription. *Computer Music Journal*, 24(2):60–76, 2000.

[2] Ali T. Cemgil, Bert Kappen, Peter Desain, and Henkjan Honing. On tempo tracking: Tempogram representation and Kalman filtering. *Journal of New Music Research*, 28(4):259–273, 2001.

[3] J. Stephen Downie. Music Information Retrieval. *Annual Review of Information Science and Technology*, 37:295–340, 2003.

[4] Anders Friberg and Johan Sundberg. Perception of just noticeable time displacement of a tone presented in a metrical sequence at different tempos. *STL-QPSR*, 33(4):97–108, 1992.

[5] John William Gordon. *Perception of Attack Transients in Musical Tones*. Dissertation, Stanford University, Center for Computer Research in Music and Acoustics (CCRMA), Stanford, 1984.

[6] Ira J. Hirsh. Auditory Perception of Temporal Order. *Journal of the Acoustical Society of America (JASA)*, 31(6):759–767, 1959.

[7] Pierre Leveau, Laurent Daudet, and Gaël Richard. Methodology and Tools for the Evaluation of Automatic Onset Detection Algorithms in Music. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR)*, Barcelona, October 2004.

[8] Ruolun Liu, Niall Griffith, Jacqueline Walker, and Peter Murphy. Time Domain Note Average Energy Based Music Onset Detection. In *Proc. of the Stockholm Music Acoustics Conference (SMAC)*, Stockholm, August 2003.

[9] Music-IR Community. 2nd annual music information retrieval evaluation exchange. Available: http://www.music-ir.org/mirexwiki/index.php/MIREX_2005, 2005. last time checked: 2005 Apr 11th.

[10] Pierre Leveau and Laurent Daudet and Gaël Richard. Sound onset labelizer and onset labels database. Available: http://www.lam.jussieu.fr/src/Membres/Leveau/SOL/SOL.htm, 2005. last time checked: 2005 Apr 11th.

[11] Rudolf A. Rasch. Synchronization in Performed Ensemble Music. *Acustica*, 43:121–131, 1979.

[12] Bruno H. Repp. Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *Journal of the Acoustical Society of America (JASA)*, 92(5):2546–2568, 1992.

[13] Bruno H. Repp. Patterns of note onset asynchronies in expressive piano performance. *Journal of the Acoustical Society of America (JASA)*, 100(6):3917–3932, 1996.

[14] Christoph Reuter. *Der Einschwingvorgang nichtperkussiver Musikinstrumente*. Peter Lang, Frankfurt, 1995.

[15] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics. Facts and Models*. Springer, 2 edition, 1999.