METHODOLOGY AND TOOLS FOR THE EVALUATION OF AUTOMATIC ONSET DETECTION ALGORITHMS IN MUSIC

Pierre LEVEAU, Laurent DAUDET
Laboratoire d'Acoustique Musicale
11, rue de Lourmel
75015 Paris - FRANCE
leveau,daudet@lam.jussieu.fr

Gaël RICHARD
GET - ENST (Télécom Paris)
46, rue Barrault
75634 Paris Cedex 13 - FRANCE
gael.richard@enst.fr

ABSTRACT

This paper addresses the problem of the performance evaluation of algorithms for the automatic detection of note onsets in music signals. Our experiments show that creating a database of reference files with reliable human-annotated onset times is a complex task, since its subjective part cannot be neglected. This work provides a methodology to construct such a database. With the use of a carefully designed software tool, called SOL (Sound Onset Labellizer), we can obtain a set of reference onset times that are cross-validated amongst different expert listeners. We show that the mean error of annotated times across test subjects is very much signal-dependent. This value can be used, when evaluating automatic labelling, as an indication of the relevant tolerance window. The SOL annotation software is to be released freely for research purposes. Our test library, 17 short sequences containing about 750 onsets, comes from copyright-free music or from the public RWC database. The corresponding validated onset labels are also freely distributed, and are intended to form the starting point for the definition of a reliable benchmark.

1. INTRODUCTION

An increasing number of studies are concerned with the automatic extraction of note onset times directly from recorded audio, as this is useful in a wide range of signal processing applications : automatic transcription, adaptive audio effects, object-based coding, and more generally all information extraction techniques used for MIR (Music Information Retrieval). All these applications try to split the audio into segments that have homogeneous properties, e.g. spectral and / or statistical properties (see for example [1, 2, 3, 4]) While this task is rather straightforward in the case of isolated notes, this can become a very difficult - and indeed ill-posed - problem for increasingly complicated sound files, from a single instrument melodic line to a full polyphonic orchestra. When many notes are played together, the notion of sound object may appear more relevant: for instance a chord can be considered as a single sound object. However, when this chord becomes broken (typically in a guitar slam) or when does it stop being a single object and start being a set of harmonically related notes ?

So far, the great majority of note onset detection schemes are based on the concept of "detection function" (DF). The DF is a highly sub-sampled version of the audio that exhibits peaks at the time instants where some properties change (e.g. energy, spectral content, etc ...) (see [5] for a tutorial on onset detection). The performance of such schemes is usually evaluated through ROC curves (Receiver Operating Characteristics), a plot of the ratio of correct detections as a function of false alarms. The main problem arises from the definition of what a "correct detection" is, since it implies the existence of a reference that gives the time localization of "true onsets" with infinite precision. Unfortunately, such perfect reference does not exist, except in a very limited set of cases (e.g. synthesized music). Furthermore, one has to allow for the finite time resolution of the above-mentioned detection algorithms : a given onset candidate at time t is counted as correct if there exists a "true onset" within a time frame $[t - \tau, t + \tau]$. Finally, the performances of the different schemes proposed in the litterature are not easily compared due to the lack of common database and protocol for their evaluation.

This paper hence addresses the two fundamental (but previously under-considered) following issues: how to build a set of reference onset times, and what is a good choice for the time resolution τ. In most cases found in the literature, the set of reference onset times is given by human-annotated data. Amongst our findings, we have observed that, for a number of test files, this human annotation exhibits a significant dependency on the employed method, the underlying software, the listener himself, and above all on the type of music. This observation suggests that the reported performance of automatic onset detection schemes is at best over-simplified and at worst cannot be generalized (i.e. are only true with strictly the same experimental conditions).

The main objective of this paper is a proposal for a common methodology and a common annotation tool, which in turn is used to build a common database of onset-annotated files. These tools and files are freely available in order to be shared by the widest community.

2. ONSET CARACTERIZATION

2.1. Particularities of onsets in music signals

Before labelling the onsets in music signals, we must define what an onset precisely is. The commonly used definition is *the time when a note begins*. However such a definition does not remove all the ambiguities. First, all the studied music signals are recorded. That implies that the *real onset time*, when the player triggers the production of the note, is not necessarily visible/audible in the signal on which we work. However we will afterwards consider that an onset is the first detectable part of a note event in the recording *if the note were isolated*. Moreover, some unwanted or uncontrolled sound events may occur when music is recorded. For instance, the keys of the woodwind and the breathing of the player produce noises that we can hear if we pay attention to it, but they usually bear little aesthetic or musical meaning. Hence, when someone is asked to label onsets in a music signal, it is important to tell him if he must take into account these events.

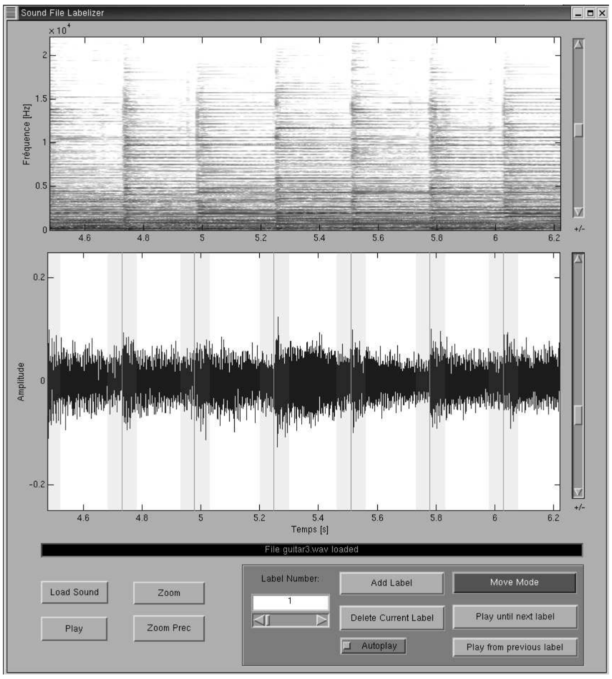
As mentioned in the introduction, picking out onsets when the notes are isolated is easy. Things begin to be more difficult when a musical sequence is played, e.g. in solo performances. For monophonic instruments, room effects are amongst the phenomena that disturb the decision, as the increased release time of a note can mask the onset of the following one. Polyphony adds other disturbances to our task: the broken chord can be considered as a sequence of notes or as a block. For bowed strings, it is also difficult to mark the onset of a note when the previous note is still played on another string. For mixed music, these difficulties are amplified. Even if the instruments are supposed to play together on a quantized temporal grid, most of the time the differences between the real onsets of the different instruments notes are not negligible, especially for slow tempi. All these elements suggest that onset detection is a relatively subjective task, and that the specifications on what we are looking for must be precisely expressed.

2.2. How to label an onset *by hand* (and by ear)?

Hand-labeling onsets is a strenuous task, that takes time and requires extreme concentration. To label onsets in a music signal, a subject can principally use three methods:

- *signal plot*: this tool is very efficient to precisely and quickly label percussive signals. It can also be used as a secondary method: when an onset occurs, the wave shape can be altered.

- *spectrogram*: it can be used as a first approach. Because of the need to take large enough FFT windows to have a sufficient frequency resolution, this method is not very precise, but it helps to localize most onsets globally. Indeed, a common characterization of music onset is that they are generally accompanied by a burst at all frequencies.



File guitar3.wav loaded

Label Number: 1

Figure 1. Interface of the *Sound Onset Labellizer*

- *listening to signal slices*: this method is the ultimate judge. Combined with visualizations, it allows an efficient labelling of signals; this is the most precise user-controlled method.

It is possible to imagine other representations of the signal, e.g. using wavelets, phase, or spectrogram scaled in bark. However, by sake of simplicity we have chosen to restrict the study to these three most commonly used methods ; and we have compiled these in a software tool called the *Sound Onset Labellizer* that is presented in next section.

3. HAND LABELLING

3.1. Annotation Tool: *Sound Onset Labellizer*

This tool has been developed to provide an easy-to-use and portable interface to the different labelling subjects. All annotators (or subjects) have used the same software.

The screen of the GUI is divided into three parts: the upper one represents the spectrogram of the signal, the middle one its time domain waveform, and the lower the controls to manipulate sound files, labels and visualization windows (see Figure 1).

The spectrogram and waveform parts have the same time axis, and all the zoom operations act on both windows. The cursor on the right of the spectrogram allows a setting of the contrast of the spectrogram, the one on the right of the signal waveform allows an amplitude magnification. The subject can play the sound visualized in the current window. The labels are put with a cursor, and can be moved by steps of 5ms to fit precisely the supposed onset time. Once a few labels are put, the subject can play the signal between two labels to evaluate if it contains only one note.

After a short learning time, all three annotators (the authors, considered as expert listeners) have spontaneously adopted similar methods to label the onsets, following these steps:

1. *zoom* on a window containing a few notes (typically 1 or 2 seconds).

2. *label with a low precision* with the help of the spectrogram.

3. *precise adjustment* with the 'autoplay' option. It allows setting the label just before a new sound event occurs.

Note that no instruction nor guidance was given before the annotation except for the tool manipulation itself.

4. TESTS

4.1. Database contents

The labelling is evaluated on a first set of 17 sound files. Most of them have been extracted from the RWC database [6]. The sampling rate used throughout is 44.1 kHz.

The other ones comes from anechoic recordings made in the laboratory. The contents of our set are shared on a web site ([7]). The sounds extracted from the RWC database are not freely available, but the references of the files are indicated, as well as the start and end samples. The self-made recordings will be shared with a free access and all the rights of use for research purposes. The completion of our database is in progress and will be updated on the web site.

The set is composed of solo performances of monophonic instruments (e.g. trumpet, clarinet, saxophone, synthetic bass), polyphonic instruments (e.g. cello, violin, distorted and steel guitar, piano) and complex mixes in different music genres (e.g. rock, classical, pop, techno, jazz).

4.2. Evaluation Methods for the annotation

In a first step, we compare the annotation of the subjects by pair. For each subject, the detected labels are counted on each file. Then, for each file, we calculate the time differences between corresponding labels where both subjects marked a given onset. The mean of these differences reveals the difficulty to annotate one file. Nevertheless this second evaluation requires an arbitrary choice: we must decide to which extent two labels must be assigned to a same onset. We have set the maximum time difference between two corresponding labels at 0.1 second, considering that it represents an upper bound for the difference in both annotators' estimates of the same onset time. However, the optimal choice of this tolerance time needs further investigations.

#	Content	Ref.	duration
1	Solo trumpet	ENST	14s
2	Solo clarinet	ENST	30s
3	Solo saxophone	ENST	12s
4	Solo synthetic bass	RWC	7s
5	Solo cello	RWC	14s
6	Solo violin	RWC	15s
7	Solo distorted guitar	6s	
8	Solo steel guitar	RWC	15s
9	Solo electric guitar	RWC	15s
10	Solo piano	RWC	15s
11	techno	RWC	6s
12	rock	RWC	15s
13	jazz (octet)	RWC	14s
14	jazz (contrabass)	RWC	11s
15	classic 1	RWC	20s
16	classic 2	RWC	14s
17	pop1	RWC	15s

Table 1. Description of our database. The files with RWC reference are taken from the RWC database, those with ENST reference are recordings made in the lab and are available on our web site. Files are grouped in 3 categories: solo monophonic instruments, solo polyphonic instruments and complex mix. Full references of the RWC files can be found on the project's web site [7]

To know the most reliable labels, we browse all the consistent labels of one comparison, and check that they are also consistent for the comparisons between the other pairs of annotations. For instance, in our case where the annotation were conducted by three subjects, the consistent labels of the comparison between subjects 1 and 2 are selected and then it is checked that they are consistent in the comparison between subjects 2 and 3, and finally between subjects 3 and 1. By computing the average times of these labels between all the annotators, reliable onset times can be obtained. It is also possible to keep only the labels of the *best labeller* (the annotator whose labels times are the closest to these average label times).

4.3. Results

The number of labels set by each user for each file, the number of reliable labels and the mean of the differences between each annotation are shown in Table 2. We can first observe that the number of labels detected by the subjects is more variable when the number of notes playable at the same moment increases. An remarkable exception is techno music: the time is so quantized that all the listeners agree to the onset repartition. Some differences also appear within the onset numbers labeled by

File #	Number of labelled onsets			Number of consistent onsets	Average timing difference
	1	2	3		
1	60	61	60	60	3.9 ms
2	38	38	46	33	13.6 ms
3	10	9	13	6	11.9 ms
4	25	25	26	25	2.5 ms
5	65	65	65	58	14.4 ms
6	79	79	79	78	7.2 ms
7	20	22	21	20	8.9 ms
8	58	58	58	58	7.7 ms
9	41	39	41	37	9.9 ms
10	20	20	20	19	7.0 ms
11	56	56	56	56	4.7 ms
12	62	62	66	59	9.9 ms
13	56	52	56	47	11.7 ms
14	61	54	52	53	9.0 ms
15	49	49	53	38	15.8 ms
16	12	12	12	4	28.4 ms
17	32	40	41	27	11.7 ms
Total	744	741	765	678	10.5 ms

Table 2. Results of the hand-labelling process. Columns marked 1, 2 and 3 represent the number of onsets labelled by each of the test listeners. The next column indicates the number of consistent onsets across listeners, used to construct our database of reliable onset times. The last column gives the mean timing error across listeners on the reliable onsets.

each user in the monophonic performances: the annotations are subjective when the breathing or the instrument keys can be heared. It emphasize the importance of precise orders to give to the listeners. If some difficulties are combined (e.g. slow attack instruments and polyphonic performance), the results cannot be exploited. For instance, only 4 onsets have been found reliable out of the 12 onsets labeled by each subject in the “classic2” file (number 16). This shows that obtaining statistically meaningful results is very complex for this file. The low number of reliable onsets is of course correlated with a high mean difference between the labels time of each subject.

5. CONCLUSION

In this paper, a fundamental aspect of the evaluation of automatic onset detection algorithms is studied. We have shown that the number of onsets detected by a listener is not only dependent on the music signal itself, but also on the guidance instructions given to annotators to mark the note onsets. This dependance suggests that onset detection algorithms could be evaluated with a different tolerance window for each type of file. For example, a 20 milliseconds tolerance window appears to be acceptable for percussive signals, while it is definitely too short for music played by bow strings to take into account the differences of onset time annotation between subjects. However, this consideration must be also balanced in regards to the application of the automatic detection. For example, if tempo detection does not demand a very accurate onset localization, the estimation of the attack duration (for example for instrument recognition) would need a far more robust onset detection function.

Finally, in order to clearly contribute to future meaningful evaluation of onset detection algorithms, the test database, the software tool used to annotate the note onsets, and the set of reliable onset times are freely available for research purposes (except for the audio files extracted from the public database RWC for which only the position of the used audio segments are provided). The perspective of evolution of the database is to include more anechoic recording of solo performances, and also to be more reliably annotated by performing further statistical analysis with a larger number of listeners.

6. REFERENCES

[1] Klapuri, A. ”Sound Onset Detection by Applying Psychoacoustic Knowledge”, *Proceedings IEEE Int. Conf. Acoustics Speech and Sig. Proc. (ICASSP)*, pp. 3089–3092, Phoenix AR, USA March 1999.

[2] Davy M. and Godsill S., ”Detection of abrupt spectral changes using Support Vector Machines: an application to audio signal segmentation”, in Proc of the IEEE-ICASSP, Orlando, Florida, 2002

[3] Goto M. and Muraoka Y., ”A real-time beat tracking system for audio signals”, Proc. of International Computer Music Conference, 1995.

[4] Rodet X. and Jaillet F. ”Detection and modeling of fast attack transients” in Proc. of International Computer Music Conference, 2001.

[5] Bello, J.P, Daudet L., Abdallah S., Duxbury C., Davies M. and Sandler M., ”A tutorial on onset detection in music signals”, *to be published (IEEE trans. on ASSP).*

[6] Goto M., RWC music database, published at http://staff.aist.go.jp/m.goto/RWC-MDB/

[7] Leveau P., Daudet L., G. Richard, ”Database and tools for onset detection evaluation” to be accessible at http://www.enst.fr/˜grichard/ISMIR04/

[8] Kauppinen I., ”Methods for detecting impulsive noise in speech and audio signals”, in Proc. of DSP-2002, July 2002.