ASSESSMENT OF STATE-OF-THE-ART METER ANALYSIS SYSTEMS WITH AN EXTENDED METER DESCRIPTION MODEL

Matthias Varewyck, Jean-Pierre Martens
Department of Electronics and Information Systems
Ghent University (Belgium)

ABSTRACT

An extended meter description model capturing the hierarchical metrical structure of Western music is proposed. The model is applied for the quantitative evaluation of four state-of-the-art automatic meter analysis algorithms of musical audio. Evaluation results suggest that the best beat trackers reach a reasonable level of performance, but that none of the tested algorithms has the potential to perform a reliable bar onset tracking. Moreover, the front-ends of the best over-all systems not necessarily seem to have the front-ends best encoding the time signature in their output. Therefore, further improvements of these systems should be attainable by a better combination of ideas that can be borrowed from existing algorithms.

1 INTRODUCTION

The temporal characteristics of music reside in an ensemble of perceivable periodic patterns at different time scales. These can be captured in a hierarchical metrical structure, called the meter [6]. The automatic extraction of metrical characteristics in musical audio is of direct importance to applications such as intelligent synchronization, standard editing, semi-automatic mixing and synchronizing audio effects but also higher-order applications such as chord recognition, structure detection and genre classification.

Western music usually exhibits a prominent periodicity, called the beat. Many meter analysis algorithms try to track these beats or to determine the corresponding tempo (e.g. [1, 2, 3]). However, tapping experiments have demonstrated [9] that the beat level is a subjective concept. Consequently, focusing too much on the beat level, may not be such a good idea. Performing an analysis involving multiple levels therefore seems a better approach. Although a number of studies describe meter analysis on symbolic (e.g. MIDI, score) data, those based on musical audio remain rather limited. In this study, we focus on the latter.

Goto & Muraoka [5, 11] considered a binary meter model with a bar, beat and intermediate level. Obviously, this model only works well if the Inter Timestamp Interval (ITI) between successive timestamps on one level is equal to two times the ITI on the lower level. Klapuri et al. [6] proposed a meter analysis involving three other levels: the bar, beat and tatum level, with ITI ratios of one to nine between subsequent levels. Klapuri's method can be applied to music with non-binary meters, but it may produce ambiguous irregular time signatures.

In this paper, we propose an extended meter description model offering a more complete representation of the time signature of a polyphonic audio excerpt, hereafter called a song. The generality of the model was first inspected by creating meter annotated data. Now, these annotations are used for the quantitative assessment of four state-of-the-art automatic meter analysis systems.

The meter description model is introduced in section 2. Section 3 explains how the annotations enabled us to design a flexible method for the quantitative evaluation of an automatic meter analysis system. Section 4 introduces the song collection and the algorithms under test (AUT's). The experiments are described in section 5, while the most important conclusions are summarized in section 6.

2 METER DESCRIPTION MODEL

The basic hypotheses of the meter description model are (1) that there always exists a quasi periodic pattern marked by main temporal accents with a mean ITI between 1.5 and 6s, (2) that secondary accents characterize the perceivable periodic patterns at smaller time scales, and (3) that these secondary accents can be represented by timestamps positioned on a uniform grid. Main accents are associated with bar transitions. Tempo changes are restricted to changes in bar lengths.

The annotation starts by tapping along with bar onsets, and continues with the manual selection of the most appropriate metrical pattern for each bar. As in [7], the model imposes constraints on the relations between successive timestamps on the same level and timestamps on successive levels. If the highest level represents the largest ITI and if it is labeled with the lowest index zero, then the main constraints are: (1) the timestamps on level L_λ are copied to level $L_{\lambda+1}$ and (2) extra timestamps on level $L_{\lambda+1}$ are obtained from those on level L_λ by dividing each ITI on L_λ into 2 or 3 equal parts. During each bar, this division ratio is the same for all ITIs on L_λ. However, for **one** L_λ, we allow that both ratios occur causing two different ITIs on L_λ.

Furthermore, per bar, a salience is assigned to each level to represent the attentional strength of the temporal

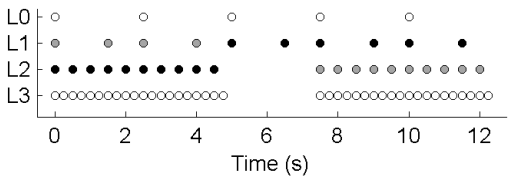


Figure 1. Example of a 3-bar annotated metrical pattern. The darkness encodes the level's salience.

pattern evoked on that level. Major and medium saliences are distinguished. One level of each bar must get a major salience while the level just above and/or below can recieve a medium salience. (see Figure 1).

3 EVALUATION METHODOLOGY

We aim to compare a sequence of hypothesized timestamps H_m ($m = 1 \dots M$) against a sequence of annotated timestamps A_n ($n = 1 \dots N$) and to retrieve a quantitative performance measure. The annotated timestamps constitute a subset of the multi-level annotation. Some interesting selection schemes are described in the next subsection.

For each A_n the set of H_m falling within a given tolerance of A_n is determined. The remaining H_m are considered insertions. If the set is not empty, H_m closest to A_n is considered as the one generated for A_n whereas the others join the insertions. This simple approach presumes that the tolerance is sufficiently small to ascertain that no hypothesis can be closer than the tolerance to more than one annotated timestamp. Given the number of annotated timestamps (N), hypotheses (M), deleted timestamps (D) and inserted hypotheses (I) the performance measures Recall (R), Precision (P) and F-measure (F) can be computed.

$$R = \frac{N - D}{N}, \quad P = \frac{M - I}{M}, \quad F = \frac{2RP}{P + R} \qquad (1)$$

3.1 Two types of experiments

In a first type of experiments we compare the hypothesized timestamps with the annotated timestamps of one single annotated level across the complete song. One important level is the *beat level*, defined as the annotation level exhibiting the largest average salience. Another important level is the *best tracked level*, defined as the level that is best supported by the hypothesized timestamps. For each AUT it is the level offering the best compromise between a high F and a good R/P balance, i.e. the level minimizing the criterion

$$E = \frac{4}{F} + \left| \frac{1}{P} - \frac{1}{R} \right| \qquad (2)$$

In a second type of experiments the hypothesized timestamps are compared with the annotated timestamps considered to represent the *correct timestamp sequence*. In case of beat tracking the *correct beat sequence* is assumed to be the annotated timestamps at the most salient level of each bar, i.e. not all the correct timestamps have to belong to the same level across the song.

3.2 Tolerance selection

Previous assessments of temporal analysis algorithms were made using a tolerance of 17.5% [5, 6] or 10% ([6], bar level). A scientific basis for the selection of a good threshold can be found in [4]. In that study, humans were asked to insert a missing fourth event in an isochronic sequence of length six. Deviations between the obtained and the mathematically correct position were about 2.5% of the ITI, with an absolute minimum of 6 ms for small ITI's. In order to adhere to other meter analysis evaluations, we use a threshold of 12.5% (5 times 2.5%) of the mean ITI, with a minimum of 30 ms (5 times 6 ms). The condition that a hypothesis cannot be within the tolerance of two successive A_n is then met if the ITI is larger than 60 ms.

3.3 Time lag compensation

Several mechanisms may result in a small but consistent time lag between the hypothesized and the annotated timestamps. This time lag may not just be punished. It was found experimentally that for a given AUT, the average time lag per song exhibits a distribution with a mean D_o that is characteristic for the AUT and with a spread that can partially originate from the different characteristics of different onsets.

Therefore, for a certain AUT, we first determine D_o as the mean time lag across a collection of songs. Per song, this value corresponds to the maximum cross-correlation between the hypothesized and the smeared annotated timestamps on the beat level. D_o is bounded between ± 100ms. The song dependent time lag $D \in (D_o\text{-}20\text{ms}, D_o\text{+}20\text{ms})$ is then determined by computing the cross-correlation a second time and finally D is subtracted from the hypothesized timestamps before supplying these timestamps to the evaluation program. The smearing converts each annotated timestamps to a gaussian with a spread equal to the time tolerance on the corresponding level.

4 EXPERIMENTAL FRAMEWORK

The evaluation set contains 30s excerpts of 161 different songs: 120 excerpts were formerly used in the MIREX 2006 beat tracking and tempo detection contests [8]. The remaining 41 were used in a human tapping experiment [9]. The average number of annotated levels per excerpt is 4.5, with a minimum of 2 and a maximum of 6. Both excerpts and annotations can be downloaded from [12].

Four algorithms, which are believed to represent the state of the art in meter analysis, were tested: one full meter analysis algorithm (KLAPURI, [6]) and three beat tracking algorithms (DAVIES [1], DIXON [2] & ELLIS [3]). All algorithms consist of a front-end producing one or more accent functions highlighting the most important rhythmic events, and a timestamp inducing back-end.

5 EXPERIMENTAL RESULTS

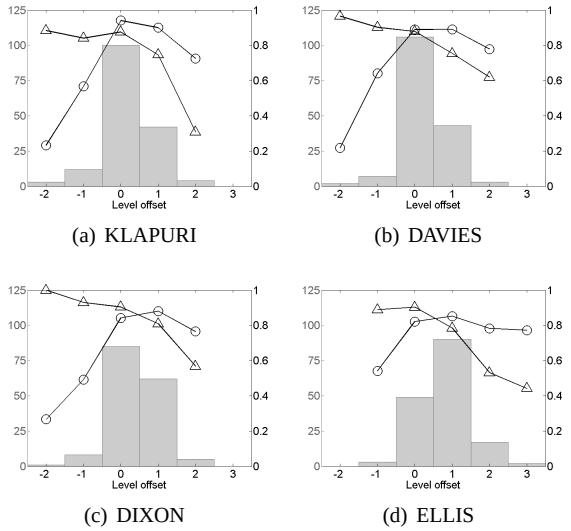
5.1 Assessing beat tracking abilities

In a first experiment we examined the beat tracking abilities of the four algorithms. In Table 1, we have listed the Precision, Recall and F-measure with respect to the best tracked level per song, as well as the number of times the best tracked level coincides with the beat level of the song. It appears that there are significant differences among the

AUT	D_o(ms)	P	R	F	N_{beat}
KLAPURI	0	0.89	0.80	0.843	100
DAVIES	-8	0.87	0.83	0.850	106
DIXON	-20	0.83	0.85	0.841	85
ELLIS	20	0.83	0.76	0.793	49

Table 1. Evaluation of beat trackers: intrinsic delay (D_o), precision (P), recall (R) & F-measure (F) w.r.t. the best tracked level, and number of songs (out of 161) for which the best tracked level is the beat level.

AUT's. Note for instance that although the F-measure of DIXON and KLAPURI are quite similar, the former only tracks the beat level in about 50% of the songs. Figure 2



(a) KLAPURI (b) DAVIES

(c) DIXON (d) ELLIS

Figure 2. Evaluation of beat trackers w.r.t. the complete metrical structure: nr. of times the best tracked level has a given offset w.r.t. the beat level, and Recall (△) & Precision (o) in the corresponding song collections.

shows that ELLIS, but also DIXON to some extent, tend to track the level below the beat level. If this happens, the Recall and Precision stay at a high level, meaning that the best tracked level is well tracked. If a level above the beat level is selected, the Precision is usually low, indicating that a large number of hypotheses did not coincide with an annotated timestamp. To complete our analysis, we measured the performances of the AUT's compared to the correct beat sequence (defined in 3.1). Results are in columns 2-4 of Table 2. Comparison with Table 1 learns

	BEAT TRACKER			FE + STIU		
AUT	P	R	F	P	R	F
KLAPURI	0.72	0.75	0.731	0.48	0.63	0.534
DAVIES	0.70	0.79	0.742	0.48	0.68	0.564
DIXON	0.62	0.82	0.709	0.42	0.51	0.463
ELLIS	0.48	0.78	0.595	0.45	0.65	0.533

Table 2. Evaluation of the full beat tracker and the front-end (FE) + simple timestamp induction unit (STIU) w.r.t. to the correct beat sequence.

that there is mainly a drop in precision which is according to the tendency of the tested beat trackers to track the level just below the beat level (see figure 2).

5.2 Role of front-end and back-end

In a second experiment, we tried to assess the role of the front-ends and back-ends as individual components. Therefore, we implemented a simple timestamp induction unit to extract timestamps directly from the accent function, without the inclusion of extra musical knowledge.

The simple timestamp induction unit (STIU) involves a peak generation and peak rejection step. The first step uses a robust left-to-right peak-valley search algorithm [10]. This first step is controlled by two parameters: the minimal time difference ΔT between successive peaks and the minimal relative drop in amplitude δ_A that must be exceeded before searching for the next peak. The peak rejection step retains all peaks exceeding some amplitude threshold TH. Per song, the STIU is run for four values of ΔT (20, 30, 40, 50 ms), three values of δ_A (0.1, 0.2, 0.3) and a user definable number of values (N_{TH}) of TH.

The performance of the STIU operated with its best (Δ_T, δ_A, TH) combination was assessed. The results are in columns 5-7 of Table 2. Comparison to the corresponding figures for the full algorithm (columns 2-4) let us conclude that the back-ends generally do a good job. Only the ELLIS back-end is not so superior to the STIU. Since the figures in column 7 of Table 2 all represent results obtained with the same back-end, they expose differences in quality of the accent functions. They suggest that the accent functions of KLAPURI, DAVIES and ELLIS carry a comparable amount of beat information. DIXON's front-end seems to be inferior to the others in this respect.

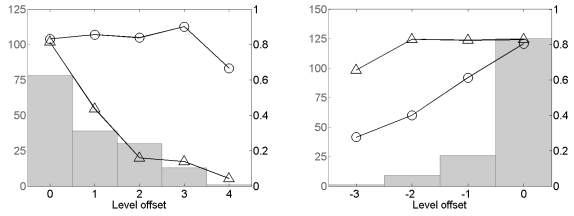
5.3 Assessing bar and tatum tracking abilities

In a third experiment we assessed the ability of KLAPURI to track the bar and tatum timestamps found on the top and bottom level in the individual bars respectively. Table 3

Tracking level	D_o(ms)	P	R	F	M/N
Bar	5.8	0.45	0.45	0.451	0.997
Tatum	0	0.75	0.70	0.725	0.933

Table 3. Evaluation of KLAPURI's bar and tatum output (M/N = nr. of hypothesized versus annotated timestamps)

shows that only 45% of the bar onsets are detected. The tatum detection results are similar to those for beat tracking. As for the *beat level*, we measured the position of the best tracked level of the bar and tatum outputs with respect to the *bar level* (the highest level in the hierarchy) and the *tatum level* (the lowest level in more than 50% of the bars). Figure 3 illustrates the results. The best



(a) Bar (b) Tatum

Figure 3. Evaluation of bar and tatum tracking: nr. of times the best tracked level has a given offset w.r.t. the bar/tatum level, and Recall (△) & Precision (o) in the corresponding song collections.

tracked level for the bar and tatum output is correct for 79 and 125 songs respectively. The bar tracking results can be explained as follows: the number of hypotheses is more or less correct (see last column in Table 3), but a large part of these hypotheses do not occur in the vicinity of a bar onset. However, they do occur very often in the vicinity of a timestamp on a lower level. By selecting that lower level, the evaluation program can significantly raise its Precision. The Recall on its turn is then expected to approach the ratio between the number of hypothesized timestamps and the number of annotated timestamps on that lower level. In case of a binary meter, the Recall at level 1 would then be close to 0.5 (we find 0.43).

For completeness, we also tested the ability of the STIU to retrieve the bar onsets from the front-end outputs. Therefore, TH is gradually increased. For a considerable number of songs, level 2 continues to be the best tracked level. This demonstrates that bar onsets are often not marked by strong peaks in the front-end output. The ELLIS front-end seems the only one producing salient peaks at bar onsets.

6 CONCLUSIONS

Four state-of-the-art beat tracking algorithms were investigated. Answers were provided to the following questions: (1) Which metric level of the song agrees best with the outputs of the algorithm? (2) How well do the outputs agree with the annotated timestamps occurring on the most salient level of each individual bar? (3) How much is the beat tracking performance affected by the front- and back-end of the algorithm? The experimental results showed that most beat trackers have a preference for tracking the beat level of the song, but for a significant number of songs they track the level just below the beat level. Between 75 and 82% of the annotated beats are correctly detected while 28 to 52% of the hypothesized beats are false alarms. The investigation of the bar and tatum tracking abilities shows that especially bar onset tracking is inadequate to serve as a basis for high-order applications. All back-ends clearly outperform a simple beat induction unit. However, we found no evidence that the incorporation of inter-level dependencies in the back-end, as in Klapuri et al [6], leads to a higher performance.

7 ACKNOWLEDGMENTS

This work was done in the context of the SEMA project, funded by BOF, Ghent. Special thanks go to Anssi Klapuri and Matthew Davies for providing their source code.

8 REFERENCES

[1] M.E.P. Davies and M.D. Plumbley, “Beat tracking with a two state model” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, vol 3, pp 241-244, 2005

[2] S. Dixon, “Automatic extraction of tempo and beat from expressive performances” in *Journal of New Music Research*, vol 30, no. 1, pp 39-58, 2001

[3] D.P.W. Ellis and G.E. Poliner, “Identifying ’cover songs’ with beat-synchronous chroma features”, in *MIREX audio cover song evaluation*, 2006

[4] A. Friberg and J. Sundberg, “Time discrimination in a monotonic, isochronous sequence”, in *Journal of the Ac. Soc. of Am.*, vol 98, no. 5, pp 2524-2531, 1995

[5] M. Goto and Y. Muraoka, “Issues in evaluating beat tracking systems”, in *Proceedings of IJCAI-97 Workshop on Issues in AI and Music*, pp 9-16, 1997

[6] A.P. Klapuri, A.J. Eronen and J.T. Astola, “Analysis of the meter of acoustic musical signals”, in *IEEE Transactions Speech and Audio Proc.*, vol 14, no. 1, 2006

[7] F. Lerdahl & R. Jackendoff, “A Generative Theory of Tonal Music”, The MIT press, 1982

[8] D. Moelants and M.F. McKinney, “Tempo perception and musical content - what makes a piece fast, slow or temporally ambiguous,” in *8th intern. conference on music perception & cognition*, Evanston, IL, 2004

[9] M. McKinney and D. Moelants,“Ambiguity in tempo perception: what draws listeners to different metrical levels?”, Music Perception, 24(2), pp 155-166, 2006.

[10] A. Vorstermans, J.P. Martens and B. Van Coile, “Automatic segmentation and labeling of multi-lingual speech data”, Speech Comm. 19, pp 271-294, 1996

[11] M. Goto, “An audio-based real-time beat tracking system for music with or without drum-sounds”, Journal of New Music Research 30(2), pp 159-171, 2001.

[12] https://speech.elis.ugent.be/ (see downloads)