Beyond the glass ceiling: what makes beat tracking difficult?

André Holzapfel* and Matthew Davies and Jose Zapata and Joao Oliveira and Fabien Gouyon

***Abstract*—abstract**

I. Introduction

Intro

II. Finding difficult songs for beat tracking

A. Evaluation Methods

Describe the available measures here and give some motivation for choosing the information gain.

B. Assigning difficulty using mutual agreement

Describe the concept of measuring the mutual agreement between beat tracking algorithms. Explain the problem of missing ground truth. Show results on D1.

C. Choice of committee members

Initially, the comittee of beat trackers contains 16 algorithms, listed in Section A. Among these, there are algorithms which perform with different accuracies according to the evaluation measures, and are characterized by varying degrees of similarities regarding their computational approach. In order to reduce the overall time for computing the mutual agreement it is reasonable to choose a subset of the 16 beat trackers. All the algorithms in this subset should be characterized by good performance, but at the same time care should be taken to include approaches that complement each other. A similar approach was investigated by Gouyon *et al.* [**?**] for the induction of tempo from music samples.

Our method to choose the beat tracking algorithms starts with computing the mean ground truth performance $S_D(b^i, a)$ for $i \in \{1 \dots 16\}$ of the i-th beat tracker on the dataset D1, using the Information Gain measure. We start including the best single performing algorithm into the subset, which is the algorithm by Klapuri *et al.* [**?**] with a mean performance of 2.237. The next step is combining it with each other beat tracker, and obtaining the *oracle* performance of each beat tracker combination. The best combination is selected, and an orcale performance vector is built that contains the oracle performance for each file. Then this oracle performance is again combined with each single beat tracker that is not yet in the subset of chosen beat trackers, and again the best combination is chosen. This procedure is iteratively continued until all beat trackers are included in the subset. We can then look at the order algorithms entered the subset, and at the performance improvement that their inclusion caused. This development is depicted in Table I.

Table I
Oracle performance of beat tracker combinations

KLA	DEG	IB2	BOE	HAI	DA1	DIX	ELL
2.237	2.399	2.470	2.518	2.553	2.582	2.599	2.610

Two observations can be made from Table I. First, a saturation effect can be observed when the number of beat trackers in the subset increases. Including more than five beat trackers does not lead to further significant improvement. And, secondly, the beat tracking approaches found at the beginning of the list are following indeed very different algorithmic approaches. Both regarding the number and the order of beat trackers, the results remain the same when using the F-measure or AMLt.

Hence we chose to use five beat trackers in our comittee. However, we decided not to use the approaches by Boeck [**?**] and Hainsworth [**?**] for reasons of portability and computation time, but included the widely available approaches by Dixon [**?**] and Ellis [**?**] instead. This lead only to a decrease in performance from 2.553 to 2.524, and will enable other researchers to easily reproduce results presented in this paper.

D. Building a difficult dataset

Describe D2 and show MMA distribution of this dataset.

III. Data annotation

A. Perceptual difficulty ratings

Clarify the scale, the exact meaning of the rating, and show the distribution of the rating for the potentially difficult dataset.

B. Spontaneous Tapping

Give details for the spontaneous tapping protocol. Maybe give some mutual agreement measurement of the tapping already here.

C. Ground truth annotation

Describe the used tools, how did we exclude impossible files based on perceptual ratings, annotation of secondary beat, give the list of tags, mention cross-checking by second expert.

INESC Porto, Porto, Portugal. aholza@inescporto.pt

D. Problems encountered during the annotation process

IV. Results

Start with the plot of BT-MMA against MGP, and also give details about the distribution of the MGP to show the difficulty of the dataset when comparing with the annotated ground truth.

Apparently, MGP is low for (hopefully) all files with low MMA. Start showing that this is only half of the work: difficulty for a beat tracker does not imply difficulty for a human listener to percieve the beat. Show a figure and discuss the los correlation between BT-MMA on one side, and the perceptual ratings and the Tapping-MMA on the other side. Show that Tapping-MMA is high for those files that were rated easy.

Conclude: there are files with perceivable meter (high Tapping-MMA), where beat trackers fail. These files can be nicely spotted by plotting BT-MMA and Tapping-MMA on top of each other. Give an analysis of these files using the tags, and maybe some feature properties. These feature properties might be motivated by the tags. For example, we might see that in this area there are mainly files with soft onsets, but no expressive timing. We might think about descriptors for that and try to separate the two groups of files.

Compare causal beat trackers with spontaneous tappings: how close are causal beat trackers to human performance for difficult files? Giva an analysis of the relations between the individual tappers as well, are there systematic differences between the tappers?

V. Conclusion

Dataset will be made available.

Appendix

List of used beat tracking algorithms

Appendix

Performance of beat trackers on datasets