ON THE AUTOMATIC IDENTIFICATION OF DIFFICULT EXAMPLES FOR BEAT TRACKING: TOWARDS BUILDING NEW EVALUATION DATASETS

A. Holzapfel, M. E. P. Davies
INESC TEC
Sound and Music Computing Group
Porto, Portugal

J. R. Zapata
Universitat Pompeu Fabra
Music Technology Group
Barcelona, Spain

J. L. Oliveira, F. Gouyon
INESC TEC
Sound and Music Computing Group
Porto, Portugal

ABSTRACT

In this paper, an approach is presented that identifies music samples which are difficult for current state-of-the-art beat trackers. In order to estimate this difficulty even for examples without ground truth, a method motivated by selective sampling is applied. This method assigns a degree of difficulty to a sample based on the mutual disagreement between the output of various beat tracking systems. On a large beat annotated dataset we show that this mutual agreement is correlated with the mean performance of the beat trackers evaluated against the ground truth, and hence can be used to identify difficult examples by predicting poor beat tracking performance. Towards the aim of advancing future beat tracking systems, we demonstrate how our method can be used to form new datasets containing a high proportion of challenging music examples.

Index Terms— Beat tracking, evaluation, selective sampling

1. INTRODUCTION

Despite the continued effort over recent years to develop beat tracking algorithms, there is a lack of diversity in the types of music signals to which these systems are applied [1]. Most systems currently developed and tested use the same or at least similar datasets [2, 3]. A recent dataset used by Grosche *et al.* [5] to examine links between local properties of a composition and beat tracking performance can be considered difficult[1], but it consists of music of a very specific style. This lack of diversity in training and testing examples can lead to beat trackers being over-fitted to particular styles of music (*e.g.* rock and pop) and can hence create a *glass-ceiling* effect, whereby beat tracking systems cease to improve in performance. Without continued effort to create new datasets which contain a wide variety of different examples for beat tracking, the difficult cases on which beat trackers currently fail are treated as outliers. While the solution to this problem of a lack of data is trivial in principle, simply "acquire more data", in practice, the collection and annotation of new datasets is a complex and extremely time-consuming procedure [4].

With the eventual aim of advancing beat tracking systems and avoiding a glass-ceiling, we believe it is necessary to assess the limitations of current beat trackers in a systematic fashion. To this end, we propose a method to automatically identify examples that are difficult for the current state-of-the-art in beat tracking. We believe that by actively seeking out difficult examples from large collections (without prior need for annotation) this will lead to the advancement of future beat tracking systems which are adapted to the properties of challenging music examples.

[1] `http://nema.lis.illinois.edu/nema_out/mirex2010/results/abt/maz/summary.html`

In machine learning research, selecting the most informative samples for training classifiers is well-known. For tasks where obtaining ground truth is costly, methods for selective sampling have been proposed [6]. In this paper, we follow a systematic approach for informative sample selection motivated by the Query by Committee concept [7]. This method provides a means to add samples to training data which increases the information that can be learned compared to a random selection of new data. In order to choose a sample, the disagreement among a committee of learners is determined, and only those samples are retained which lead to a high degree of disagreement. In contrast to approaches for instance selection such as the one presented by Wilson and Martinez [8], this technique does not require prior labelling or annotation of the data. Similar concepts have been proposed in the domain of speech processing [6], but, to the best of our knowledge, this technique has not yet been applied to music signal processing applications like beat tracking.

For our application we consider a collection of beat tracking algorithms to be a committee of learners. Instead of using their outputs *e.g.* in a fusion system for beat tracking, we measure their pair-wise disagreement to indicate the degree of difficulty of a specific music sample. We show that, even in absence of ground truth, this disagreement measure can serve as a way to select new music samples for manual annotation, and that it can be used to estimate the overall difficulty of available datasets.

However, beat tracking is not a simple classification task, the concept of "beat" in music is highly subjective and there are several ways in which the outputs of beat trackers can differ while still being considered related, *e.g.* two sequences which are tapped at different metrical levels or in anti-phase to one another. Therefore, an important contribution of this work will be to choose an appropriate evaluation measure as a basis for determining disagreement; one that can contend with the ill-posed nature of beat tracking. Through a comparison of evaluation methods we demonstrate that the Information Gain method [9] is best able to identify the musical excerpts where beat trackers mutually disagree, and hence we propose its use for forming a new dataset comprised of challenging examples.

Furthermore, our experiments show that the proposed method is effective even with a small number beat trackers in the committee, provided these beat trackers are diverse and are shown to perform accurately. In this way our method can be applied where there is value in knowing if a signal to be analysed will be hard to beat track, without the need for ground truth annotations; for example when deciding whether to use beat-synchronous or fixed-time analysis frames for cover song detection [10].

The remainder of this paper is structured as follows: in Section 2 we provide a summary of methods applied to evaluate beat track-

ing systems. In Section 3 we describe our method for determining the mutual disagreement of beat tracking systems. In Section 4 we include results to demonstrate the validity of the proposed method, and compare the properties of different evaluation methods. In Section 5 we apply our method to a new dataset for which no ground truth exists, and present conclusions in Section 6.

2. BACKGROUND

When comparative studies of beat tracking algorithms have been undertaken the goal is usually two-fold, first to determine which algorithm is the most accurate when compared to the ground truth and then to verify whether any observed differences in the performance of the algorithms are statistically significant. However when performing these comparisons, the choice of evaluation method can have a critical impact on the observed outcome, potentially changing the ranking of algorithms and presence of significant differences [9]. But why should such inconsistencies exist for beat tracking evaluation?

All methods share the common aim of measuring the extent to which a meaningful relationship between the output of a beat tracking algorithm and a sequence of ground truth annotations exists. However, there is currently no consensus on how to achieve this goal. This has led to evaluation methods with differing properties. In this paper we focus on three techniques to approximately cover the range of existing methods:

- F-measure [3]: Beats are considered accurate if they fall within a $\pm$70ms tolerance window around annotations. Accuracy in a range from 0% to 100% is measured as a function of the number of true positives, false positives and false negatives.
- AMLt [4]: A continuity-based method, where beats are accurate when consecutive beats fall within tempo-dependent tolerance windows around successive annotations. Beat sequences are also accurate if the beats occur on the off-beat, or are tapped at double or half the annotated tempo. The range of values for AMLt is 0% to 100%.
- Information Gain [9]: Accuracy is determined by forming a histogram of the timing error between the beat sequence and annotations, from which a numerical score is calculated as a function of the entropy of the histogram. The range of values for the Information Gain is 0 bits to approximately 5.3 bits.

Our aim is to reliably identify difficult pieces, *i.e.* where performance of beat trackers is poor and their output has no relation with the actual beats of the piece. We are looking beyond the so-called errors which reflect ambiguity in metrical level (octave error) and beat phase (off-beat tapping) towards identifying the complete failure of the algorithms. Therefore, it is important to consider the conditions where these evaluation methods assign 0% accuracy: F-measure – when the beats and annotations are in anti-phase, or in the (unlikely) case that no beats fall within any tolerance windows; AMLt – when the metrical relationship between beats is not related by a simple factor of two, or if no beats fall within the specified tolerance windows; Information Gain – in the limit when the beat sequence and annotations are completely unrelated. For further details see [9].

3. APPROACH

Given a committee of N beat trackers B_i, $i = 1 \dots N$, and a set of beat annotated samples x_k, $k = 1 \dots K$, it is a straightforward procedure to get an estimation of the difficulty of each sample for beat tracking. First, for each beat tracker B_i a beat tracking sequence b_k^i for song x_k is obtained. Such sequences and annotations consist of a list of time instants. Given the annotation sequence a_k for excerpt x_k, and an evaluation measure, the performance $S(b_k^i, a_k)$ on x_k can be determined for each beat tracker. An objective definition of difficulty for beat tracking is simply to compute the mean ground truth performance (MGP) among all beat trackers, which we denote as $\bar{S}(b_k^i, a_k)$, where an example is considered difficult if the mean performance among beat trackers is low.

If no ground truth is available for a given example, then we cannot use MGP to infer the level of difficulty. However, given only the beat tracking outputs we can compute a mutual agreement $S(b_k^i, b_k^j)$ between the outputs of beat trackers B_i and B_j for song x_k. Following the Query by Committee concept [7], the most informative samples can be characterised by low mutual agreement. More intuitively, an unknown sample might be "interesting" for beat tracking if the committee of beat trackers disagree in their estimates of the beat. Thus, the mean of all mutual agreements (MMA) between the beat tracking estimations, $\bar{S}(b_k^i, b_k^j)$, is chosen as an indicator of an informative sample.

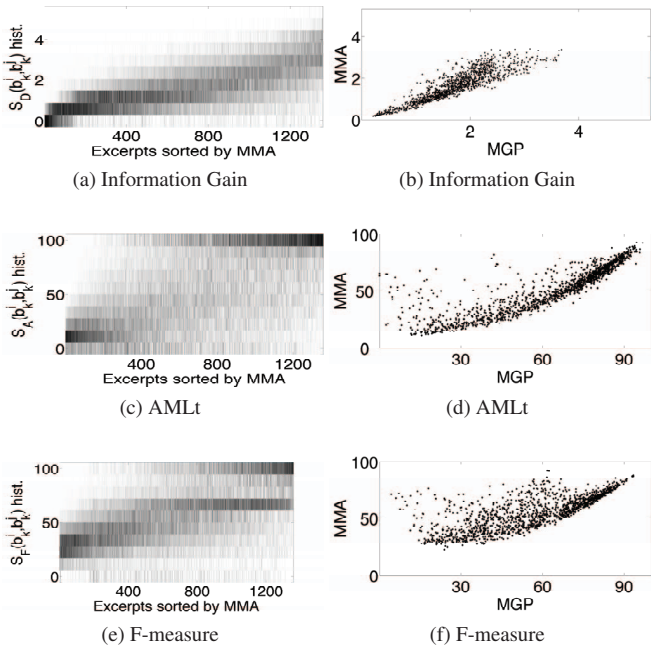
Therefore when choosing an evaluation method for this purpose we need to be sure that low values of mutual agreement are indeed indicative of unrelated beat sequences. We should be aware that, although commonly used for beat tracking evaluation, the F-measure does not fulfil this criterion, as a value of zero is most commonly the result of comparing two phase shifted pulse trains of equal period. While being able to identify this property may be of interest (*e.g.* to discover ambiguity in consistently identifying the phase of the beats), it is not in line with our goal of estimating the difficulty of musical excerpts as two equal but phase inverted beat tracking sequences cannot be said to be in complete disagreement. Furthermore, the widely used AMLt measure has the similar property that any period or phase relation between two sequences will be penalised with a zero value if it is not explicitly specified in the calculation. On the other hand, the Information Gain method is characterised by a true zero value (when the two beat sequences are completely unrelated) and has a continuous range of performance that does not require the specification of acceptable relations as with AMLt. In the following section, we will compare the effect of varying the evaluation method towards the aim of estimating beat tracking difficulty.

4. RESULTS

To validate our method we collect the output of 16 beat tracking algorithms on a large dataset formed as a superset of existing beat tracking datasets. It contains 1360 excerpts [3] and we refer to it as **Dataset 1**. Note that both MMA and MGP are computed for the whole excerpts, thus resulting in a global measure of difficulty for a whole music sample which does not take into account possible variations within a sample. Where not publicly available, the implementations of the algorithms were obtained directly from the authors; the algorithms were chosen to cover a wide variety of technically independent approaches. Note that in this paper our goal is not to undertake a comparative study of beat tracking algorithms, therefore the individual performance of the beat tracking algorithms will not be presented.

The output of each algorithm is tested using the beat tracking evaluation toolbox[1]. We limit ourselves to the three methods de-

[1] http://code.soundsoftware.ac.uk/projects/beat-evaluation



(a) Information Gain (b) Information Gain

(c) AMLt (d) AMLt

(e) F-measure (f) F-measure

Fig. 1: Left column: Histograms of the mutual agreement $S_z(b_k^i, b_k^j)$, sorted by their mean values (MMA). Dark colors indicate high histogram values. Right column: MMA versus MGP scatter plots.

scribed in Section 2: F-measure, AMLt and Information Gain which label F, A, and D respectively. For each evaluation method we record the mutual agreements $S_z(b_k^i, b_k^j)$ between each pair of beat trackers and the scores relating the beat trackers to the ground truth annotations, $S_z(b_k^i, a_k)$ where $z \in \{F, A, D\}$ represents a given evaluation method.

For each musical excerpt, we form a histogram of mutual agreement values. To visualise behaviour across the entire data set we create a matrix where each column represents a different excerpt and we sort the columns in ascending order according to the MMA for each evaluation method, $\bar{S_z}(b_k^i, b_k^j)$. The MMA sorted histograms are shown in the left column of Figure 1. As can be seen from the histogram images there is a general trend from the bottom left to the top right hand corners representing two types of behaviour: bottom left – low MMA values indicate mutual disagreement between the beat trackers; top right – high MMA values indicates similar outputs of the beat trackers.

Comparing the different evaluation methods we can explore the extent to which this pattern is consistently present. For the Information Gain method (Fig. 1a) the low agreement end of the histogram image is most pronounced (lower left corner). Conversely for AMLt, the cases where beat trackers agree with each other are clearest, as seen in the high amplitude cluster in the upper right hand corner of Fig. 1c. F-measure (Fig. 1e) is different from AMLt in two respects. First, due to the unequal treatment of metrically related sequences (which are considered equally valid for the AMLt calculation), the histograms with high MMA are bi-modal having peaks at 66% and 100%. Second, the lowest MMA region is considerably higher than zero, with an F-measure of approximately 30%. This behaviour is the result of a significant proportion of beats accidentally falling within the allocated tolerance windows without the existence of a meaningful relationship [9].

While the patterns shown in the mutual agreement histograms indicate that there is a range of agreement and disagreement between the outputs of the committee of beat trackers this information alone is not informative unless we can demonstrate that the MMA is related to the performance of the beat trackers against the ground truth. To this aim we examine scatter plots of MMA and MGP for each evaluation method, which are shown in Figures 1b, 1d & 1f.

Comparing the scatter plots for each evaluation method we can observe the following positive correlations, r_z, between MMA and MGP, $r_F = 0.74$, $r_A = 0.86$, $r_D = 0.90$. From this behaviour we can infer that when there is high agreement between beat trackers this is indicative of high mean performance of the beat trackers against the ground truth. Similarly at the other end of the scale, mutual disagreement implies low mean performance against the ground truth. While this general trend exists for each evaluation method, the shapes of the scatter plots are not identical. Perhaps the most noticeable difference is that the greatest coherence between MMA and MGP exists for Information Gain for poor performance (Fig 1b) but for AMLt and F-measure the correlation is strongest for accurate performance (Figs. 1d and 1f).

Given the high correlation between MMA and MGP and our intention to use an evaluation method to automatically find challenging examples for beat tracking, we recommend using MMA driven by the Information Gain method to predict the mean performance of the committee of beat trackers against the ground truth.

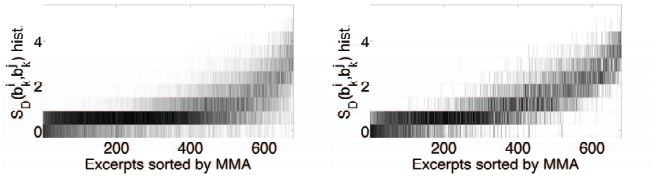
5. FORMING A NEW DATA SET

Having examined the relationship between MMA and MGP we now turn our attention towards finding interesting examples for beat tracking without the need for ground truth annotations. By running our committee of beat trackers on unseen data and deriving the MMA (using Information Gain) we can estimate the degree of difficulty for a given excerpt.

Towards this aim, we set out to create a new data set, named **Dataset 2**. It contains examples that, based on our intuition, we considered would have challenging properties for beat trackers, including: a lack of prominent percussion; changes in tempo; wide dynamic range; poor audio quality; pauses; and changes in time-signature. **Dataset 2** contains 678 excerpts of $40s$ length from various musical styles such as classical, chanson, jazz, folk and flamenco.

Since the dataset is not annotated we cannot provide a scatter plot to compare MMA with MGP, but we can examine the mutual agreement histogram image. The differences between **Dataset 1** and **Dataset 2** can be seen by comparing Figures 1a and 2a. Here we see a similar general pattern across both datasets, but **Dataset 2** has a greater proportion of musical excerpts with low MMA. Given the relationship between MGP and MMA for **Dataset 1** and the general similarity in shape, we believe that **Dataset 2** does contain a high proportion of challenging excerpts for beat tracking. However to formally verify this hypothesis we plan to annotate **Dataset 2** in our future work.

To this point all of the analysis has been based on the output of mutual agreement between 16 beat tracking algorithms. While it is informative to examine the global properties of as many beat tracking systems as possible, in practice it was a complex undertaking to collect, compile and execute these algorithms over two large datasets which required multiple operating systems and dedicated computing resources. Therefore, a more practical solution would be to see if the same behaviour holds for a smaller committee. To this aim we



(a) 16 Beat Trackers (b) 5 Beat Trackers

Fig. 2: Information Gain histograms of mutual agreement $S_D(b_k^i, b_k^j)$ sorted by MMA for **Dataset 2**. Dark colors indicate high histogram values.

generate an MMA sorted histogram image as in Figure 2a but using a subset of five beat trackers ([2, 3, 11, 12, 13]). This subset included beat trackers that performed accurately against the ground truth for **Dataset 1**, but did not have any technical dependencies and were developed in different research groups. The combination of high accuracy and diversity is important for a committee of learners [14], but due to space restrictions a systematic analysis for the choice of beat trackers is omitted here. Furthermore, all five chosen methods are either available under public licenses or can be obtained for research purposes. The five beat tracker histogram image is shown in Fig. 2b.

Comparing Figures 2a and 2b, a near-identical trend with the low MMA excerpts is clearly visible. However due to fewer pairwise comparisons between beat trackers, the data is sparser. Given this similarity in shape we infer that a smaller committee of experts can be sufficient to estimate the difficulty of music examples for beat tracking. However this inference should only apply if the beat trackers chosen to form the smaller committee are shown to perform well against some available ground truth; that is, a committee of poor performing beat trackers will likely not provide useful information about the potential difficulty of examples in a dataset, since, for a bad beat tracker, many excerpts may appear difficult.

6. CONCLUSIONS

In this paper we have a presented a technique for estimating the degree of difficulty of musical excerpts for the current state-of-the art in beat tracking based on the mutual agreement between a committee of beat tracking algorithms. We have demonstrated that the mean mutual agreement between beat tracking outputs is correlated with mean beat tracking performance against ground truth when tested using three different evaluation methods. Furthermore we showed how to use this relationship for forming new datasets with a bias towards challenging examples.

It is important to note that this behaviour can only be used to summarise the behaviour of the committee of beat trackers. We cannot make any inferences about individual beat tracking systems, indeed it is possible that one beat tracker in the committee could agree exactly with the ground truth annotations and this would not be observable. However we consider this an unlikely outcome.

It remains an open question as to how far the difficulties of beat tracking methods are related to difficulties human listeners have in tapping the beat to music. In future work we will record subjective ratings from human expert listeners to describe the perceived difficulty of tapping to the examples in **Dataset 2**. With this information we plan to explore the relationship between mutual agreement of beat trackers and perceptual difficulty, in particular we will seek to find musical excerpts which appear perceptually easy, but remain challenging for beat trackers. We believe that determining the musical and acoustic properties of such signals will be of key importance towards advancing beat tracking techniques.

7. ACKNOWLEDGEMENTS

This research received support from the Portuguese Foundation for Science and Technology through the project "ShakeIt" (grants UTAustin/CD/0052/2008 and PTDC/EAT-MMU/112255/2009) and through grants SFRH/BD/43704/2008 and SFRH/BPD/51348/2011, and by Universidad Pontificia Bolivariana (Colombia) and Colciencias, and by the EU-funded project MIReS.

8. REFERENCES

[1] N. Collins, "Towards a style-specific basis for computational beat tracking," in *Proc. of the 9th International Conference on Music Perception and Cognition*, 2006, pp. 461–467.

[2] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[3] S. Dixon, "Evaluation of the audio beat tracking system BeatRoot," *Journal of New Music Research*, vol. 36, no. 1, pp. 39–50, 2007.

[4] S. W. Hainsworth, *Techniques for the Automated Analysis of Musical Audio*, Ph.D. thesis, Cambridge University, Department of Engineering, 2004.

[5] P. Grosche, M. Meinard, and C. S. Sapp, "What makes beat tracking difficult? a case study on Chopin mazurkas," in *Proc. of the 11th ISMIR conference*, 2010, pp. 649–654.

[6] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proc. of the 12th International Conference on Machine Learning*, 1995, pp. 150–157.

[7] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. of the 5th annual workshop on Computational learning theory*, 1992, pp. 287–294.

[8] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 257–286, 2000.

[9] M.E.P. Davies, N. Degara, and M.D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Tech. Rep. C4DM-TR-09-06, QMUL, C4DM, 2009.

[10] J. P. Bello, "Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats," in *Proc. of the 8th ISMIR conference*, 2007, pp. 239–244.

[11] J. Oliveira, F. Gouyon, L. Martin, and L. Reis, "IBT: A realtime tempo and beat tracking system.," in *Proc. of the 11th ISMIR conference*, 2010, pp. 291–296.

[12] N. Degara, E. Argones, A. Pena, M. Torres, M. E. P. Davies, and M. D. Plumbley, "Reliability-informed beat tracking of musical signals," *IEEE Transactions on Audio, Speech and Language Processing, in press*, 2011.

[13] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[14] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Proceedings of the 21st International Conference on Machine Learning*, 2004, pp. 74–81.