Selective sampling for beat tracking evaluation

André Holzapfel*, Matthew E. P. Davies, José R. Zapata, João Lobato Oliveira *Student Member, IEEE,* and Fabien Gouyon *Member, IEEE*

***Abstract*—In this paper, we propose a method that can identify challenging music samples for beat tracking without ground truth. Our method, motivated by the machine learning method "selective sampling", is based on the measurement of mutual agreement between beat sequences. In calculating this mutual agreement we show the critical influence of different evaluation measures. Using our approach we demonstrate how to compile a new evaluation dataset comprised of difficult excerpts for beat tracking and examine this difficulty in the context of perceptual and musical properties. Based on tag analysis we indicate the musical properties where future advances in beat tracking research would be most profitable and where beat tracking is too difficult to be attempted. Finally, we demonstrate how our mutual agreement method can be used to improve beat tracking accuracy on large music collections.**

***Index Terms*—Beat tracking, selective sampling, evaluation, ground truth annotation.**

I. Introduction

The task of automatic extraction of beat times from music signals is a mature research topic within music information retrieval (MIR). The aim of a beat tracking system is to recover a sequence of time instants consistent with how a human might tap their foot in time to music. Used in this way beat trackers have become standard tools within other MIR problems (e.g. structural segmentation [1], chord detection [2], music similarity [3]) by enabling "beat-synchronous" analysis of music. While many different techniques have been presented for beat tracking, in particular over the last five years (*e.g.* [4], [5], [6], [7], [8], [9]), analysis of beat tracking accuracy reveals there has been little significant improvement over the method of Klapuri et al [10] from 2006 which is still widely considered to represent the state of the art. One reason for this apparent stagnation might be that beat tracking has simply reached the upper limit of performance (the so-called "glass-ceiling" effect) and no further gains in performance are possible. Perhaps a more likely explanation lies in the data used to evaluate beat trackers. We believe the continual re-use of existing datasets (e.g.[10], [11], [6]) has led to a (somewhat) inevitable over-fitting of beat tracking algorithms to the limited data which is available. Furthermore, within these existing databases, there is a bias towards musical styles considered easier for beat tracking, including: rock, pop and electronic dance - genres typically characterized by clear percussive content and steady tempi. This imbalance towards easier musical styles means that challenging excerpts, where beat tracking algorithms fail, are typically treated as outliers and little effort is made to determine how to process them.

Given the hypothesis that a glass ceiling in beat tracking exists due to a lack of diversity in annotated data, an appropriate strategy would be to annotate more musical examples. However the manual annotation of beat locations can be extremely difficult and time-consuming. Therefore it makes sense to restrict annotation to music examples which are in some way informative for the beat tracking problem. To this end our approach is to focus on the selection of musical pieces that are shown to be difficult for current state of the art systems. Since the goal is to subsequently derive ground truth annotations, this estimation of difficulty must be achieved without any ground truth annotations.

While some effort has been made to estimate rhythmic difficulty, this has typically been limited in scope focusing on measures of beat strength [12], [13]. Furthermore these methods have not been used for the selection of music samples to annotate. A related study of difficulty in beat tracking by Grosche *et al.* [14] considered local properties of compositions that cause beat trackers to stumble, whereas our interest is in the global properties of musical excerpts.

In machine learning research, selective sampling approaches have been proposed to select informative samples in absence of ground truth [15]. In this paper, we follow the Query by Committee concept [16] and assign a degree of difficulty to a given piece by measuring the mean mutual (dis-)agreement (MMA) between a set of state of the art beat tracking approaches. In effect, when there is no consensus among the beat tracking algorithms we consider that the music example in question might be difficult. When assembling our committee of beat trackers, we take into account that the committee should be characterized both by high accuracy and diversity [17]. Similar concepts have been evaluated in the domain of speech processing [15], and Mandel *et al.* [18] presented an approach which includes user interaction to identify informative samples for training a music retrieval system. However, to our knowledge, selective sampling has not yet been applied in the evaluation of music signal processing tasks like beat tracking.

While the basic concept of selective sampling for beat tracking evaluation was introduced in [19], an important aspect we consider in this paper is to what extent the musical prop-

This work is partly supported by the European Commission, FP7 (Seventh Framework Programme), ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711; the European Regional Development Fund through the Programme COMPETE; by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583); by National Funds through the Portuguese Foundation for Science and Technology, within projects ref. PTDC/EAT-MMU/112255/2009 and PTDC/EIA-CCO/111050/2009 and SFRH/BPD/51348/2011; by Universidad Pontificia Bolivariana y Colciencias.

AH was with INESC TEC for a large part of this work, and is now with Universitat Pompeu Fabra, Barcelona, Spain, e-mail: (hannover@csd.uoc.gr). MEPD is with INESC TEC, e-mail: (mdavies@inescporto.pt). JRZ is with Universitat Pompeu Fabra, Barcelona, Spain, e-mail (joser.zapata@upf.edu). JLO is with the Artificial Intelligence and Computer Science Laboratory (LIACC) at the Faculty of Engineering of University of Porto (FEUP) and with INESC TEC, e-mail: (joao.lobato.oliveira@fe.up.pt) FG is with INESC TEC and FEUP, e-mail: (fgouyon@inescporto.pt).

erties that make beat trackers fail coincide with the properties that make tapping to a piece difficult for human listeners. To this end we used the proposed MMA method to build a dataset of samples that are problematic for beat trackers. Listeners were then asked to tap the beat of those pieces in a spontaneous manner, to describe the signal properties, and eventually to determine ground truth beat annotations. This data was used to investigate similarities and differences between human listeners and automatic beat tracking. Results demonstrate that among the files shown to be difficult for beat trackers some were perceptually easy for human tappers, while those files characterized by expressive timing and/or quiet accompaniment were considered just as difficult. We believe that the highest potential for improving beat tracking technology lies in determining methods to address those files that cause beat trackers to fail but which contain a perceivable beat, rather than attempting to address those for which human tappers also struggle to infer the beat.

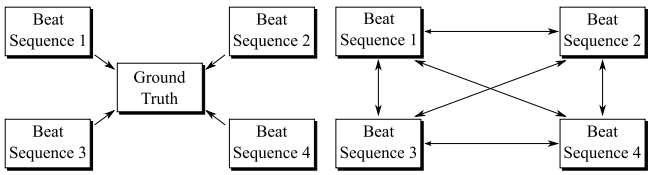
The remainder of the paper is structured as follows. In Section II, we motivate the usage of mutual agreement for detecting difficult samples and address issues of evaluation measures and the choice of beat tracking algorithms for mutual agreement computation. In Section III, we use an existing beat tracking database to determine system parameters for the MMA computation, and demonstrate the validity of our approach. In Section IV, we give details about a new dataset compiled for this publication and the annotation process. In Section V, we investigate the difficulty of the new dataset both for automatic beat tracking and human listeners. In Section VI, we describe the application of our MMA method to identify and reject musical pieces where beat tracking will fail, and furthermore demonstrate how beat tracking performance can be improved directly by inspecting the properties of the beat tracking committee. Finally, in Section VII we give a summary of the principal findings and an outlook towards future work.

II. Mutual sequence agreement

Our approach is motivated by the Query by Committee concept [16], and provides a method for selecting informative data samples to add to existing training data. While most beat tracking systems are optimized manually, we can compare this optimization process with a learning process, and the current state of the art can be considered a committee of learners that can profit from selecting informative new training samples.

A graphical representation for estimating the difficulty of a music sample for beat tracking when ground truth is given is shown in Figure 1a. Here, a set of N beat sequences is calculated for a given sample using N different beat trackers. These beat sequences are then compared with the given ground truth of the piece using an evaluation measure, and the *mean ground truth performance* of all beat trackers, **MGP**, on this piece can serve as an estimate of its difficulty. Note that this is different from calculating the mean performance of a single beat tracker over an entire data set, which can serve as an indicator of its individual performance.

However, when no ground truth is given, an unknown sample might be labeled as "interesting" for beat tracking if



(a) Ground truth given (b) no ground truth

Figure 1: Setups for determining difficulty of a sample for $N = 4$ beat trackers, (a) with and, (b) without ground truth.

a committee of beat trackers disagree in their estimates of the beat. Hence, the beat sequences of the N beat trackers are compared with each other, creating a complete graph with $N(N-1)/2$ mutual agreement values on its edges, as shown in Figure 1b. The mean weight of the edges is equal to the *mean mutual agreement* between the beat sequences, **MMA**, which we investigate as a method for estimating the beat tracking difficulty. When specifically referring to beat tracking outputs we will use the notation **BT-MGP** and **BT-MMA**.

To use this technique for beat tracking we must address two important decisions: first, which evaluation method to use to compute the mutual agreements between committee members and second which beat trackers to include in the committee.

A. Evaluation Measures

Our mutual agreement measure relies on the use of an objective beat tracking evaluation method to determine the relationship between pairs of beat sequences. The selection of this evaluation method poses an immediate problem since there is no commonly accepted technique for measuring beat tracking performance. This lack of consensus has led to many approaches being developed, each with differing parameters and/or methodologies. For a review and further discussion, see [20]. The variations among evaluation methods arise due to differing hypotheses on how to address the localization between beat times and annotations (e.g. by the use of tolerance windows), and how to contend with ambiguity over the validity of metrically related sequences. The eventual choice of a specific evaluation method is usually made in the context of a particular application. For example, when evaluating a real-time beat tracking system, a continuous relationship between beats and annotations may be an important criterion [21]. Or, for chord recognition, permitting many different interpretations of the beat may be detrimental to chord detection accuracy [22] hence it may be advisable to restrict the range of alternate interpretations of the beat.

Our motivation for using a beat tracking evaluation method is somewhat different, since our primary interest is not in identifying where beat sequences agree with each other per se, but rather in finding cases where they disagree. While this disagreement could be measured in terms of ambiguity in metrical level or beat phase, this is of limited use since these beat sequences could be considered "somehow" related. Of greater importance for our application is finding when the beat sequences are completely unrelated. This is based on our intuition that beat trackers are usually built out of similar

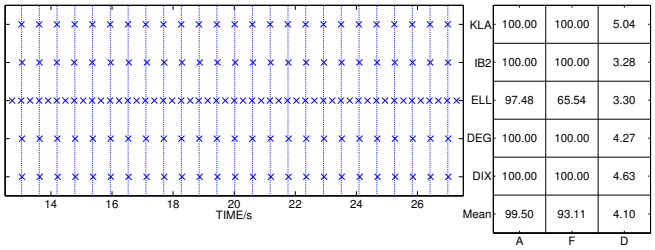
components, and therefore a significant lack of consensus in their outputs should be indicative of something interesting in the input signal. Based on this reasoning, the choice of evaluation method may appear trivial, since we could simply look for cases where the evaluation score was close to 0% for any evaluation method. To explore this hypothesis further we briefly address the properties of three evaluation methods which cover the main types of techniques currently used. For each we describe its basic functionality and indicate the conditions under which a minimal accuracy score can occur.

F-measure [6]: Beats are considered accurate if they fall within a $\pm$70ms tolerance window around annotations. Accuracy in the range 0% to 100% is measured as a function of the number of true positives, false positives and false negatives. If the beat sequences are tapped at metrical levels related by a factor of two (but otherwise well aligned), this causes the score to drop from 100% to 66.7%. A score of 0% can only occur if no beat times fall within any tolerance windows. The most likely scenario for this score is if the beat sequences tapped in anti-phase (i.e. on the "off-beat"). Completely unrelated beat sequences typically score around 25% by virtue of beats arbitrarily falling within the range of tolerance windows [20].

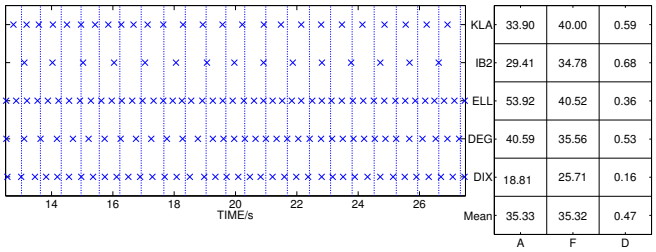
AMLt [11]: A continuity-based method, where beats are accurate when consecutive beats fall within tempo-dependent tolerance windows around successive annotations. Beat sequences are also accurate if the beats occur on the off-beat, or are tapped at double or half the annotated tempo. The range of values for AMLt is 0% to 100%. A score of 0% can only occur if no two consecutive beats fall within the specified tolerance windows. This is most likely the result of the beat sequences being related by an unspecified metrical relationship, e.g. "2 against 3" [23]. As with F-measure, unrelated sequences do not score 0%, being closer to 18% [20].

Information Gain [23]: Accuracy is determined by calculating the timing errors between an annotation and all beat estimations within a one-beat length window around the annotation. Then, a beat error histogram is formed from the resulting timing error sequence. A numerical score is derived by measuring the K-L divergence between the observed error histogram and the uniform case. This method gives a measure of how much information the beats provide about the annotations. The range of values for the Information Gain is 0 bits to approximately 5.3 bits, where the upper limit is $\log_2(K)$ for K histogram bins. Maximal Information Gain is the result of all beat error measurements falling within a single histogram bin, hence the choice of K is important and should be neither too large nor too small; K = 40 histogram bins is an appropriate choice [23]. An Information Gain of 0 bits is obtained, in the limit, when the beat error histogram is uniform, i.e. where the beat sequences are totally unrelated.

Based on properties of these evaluation methods, the Information Gain approach would appear most suited to our purpose since it is the only method guaranteed to be close to 0 only in the condition where the beat sequences have no meaningful relationship. However, to confirm this empirically we retain all three evaluation methods throughout the subsequent analysis. In our notation, we will add a subscript $z \in \{F, A, D\}$ for F-measure, AMLt and Information Gain, respectively, whenever a distinction is of importance (*e.g.* BT-MMA$_D$ for BT-MMA using Information Gain).



(a) Example for an easy song (Busta Rhymes)



(b) Example for a difficult song (Tom Waits)

Figure 2: Ground truth annotations for two songs shown as dotted vertical lines. Beat estimations for five algorithms are superimposed as crosses. The tables list the ground truth performance according to the three evaluation methods for each songs, and their mean. A and F are measured in % while D is measured in bits.

To illustrate the differences in beat tracking outputs and the effect of different evaluation methods we examine two examples. The first, in Figure 2a, shows beat estimations that strongly agree with one other. The third sequence tapped at twice the tempo, causes an expected drop in F-measure but the mean performance of all algorithms against the ground truth is very high. However in Figure 2b, there is much less agreement between the beat sequences and this is reflected in the performance against the ground truth. Despite this mutual disagreement, the mean performance of the algorithms for F-measure and AMLt is still around 35%. While the Information Gain (D) is measured on a different scale, it is much closer to its theoretical lower limit.

B. Choice of committee members

In the first phase of this research project, implementations of various beat tracking algorithms were collected including those freely available online and others kindly provided by the authors of the systems on request. In total we compiled an initial committee of 16 beat trackers listed in Table I.

In practice, this required considerable effort to install appropriate system components and operating systems necessary to make all of the algorithms run. Furthermore there was both considerable variability in the computational complexity of the algorithms, with some algorithms slower than the fastest by up to two orders of magnitude, and large variation in beat tracking performance (see Section III). Towards making the results of this paper more easily reproducible we propose a method to select a subset of these algorithms. The selected

algorithms should be characterized by good performance, but at the same time care should be taken to include approaches that complement each other. The goal is to obtain a small but diverse committee, where each implementation is publicly available and not too demanding in terms of execution time.

To find a subset of the $N = 16$ beat tracking algorithms we make use of an *oracle* method. The first stage in this method is to run all beat tracking algorithms on an existing annotated dataset recording the per track performance of each algorithm. The first member of the committee is the algorithm which performs best in the mean across the entire dataset. The next member to enter to the committee is determined by an iterative method. Each remaining algorithm is taken in turn and it is combined with those currently in the committee – in this case just the first algorithm. The oracle performance is recorded by selecting the most accurate algorithm per track in the dataset. Whichever of the remaining algorithms gives the greatest improvement in oracle performance is the next to enter the committee. This procedure is iteratively continued until all beat trackers have been included. We can then look at the order in which the algorithms entered the committee and the improvement in performance achieved by their inclusion. We can determine a subset by fixing the number of committee members at the point where improvements offered by additional members is small. A choice of beat trackers guided by this strategy takes into account both accuracy and diversity.

III. Applying MMA to an existing dataset

The largest dataset for beat tracking evaluation to date was introduced by Gouyon [24]. It contains a total of 1360 excerpts from different styles of music and will be referred to as **Dataset1** throughout this paper. We use Dataset1 to investigate the accuracy and diversity of the available 16 beat trackers. Based on these results we will i) select our committee of beat trackers ii) give a proof of concept for our MMA method to assess difficulty for automatic beat tracking and iii) determine the most appropriate evaluation method.

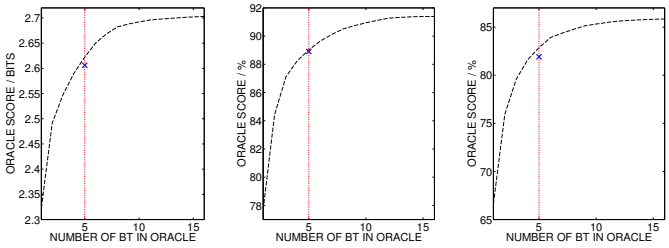
A. Accuracies of potential committee members

In Table I the individual ground truth performance of each of the 16 beat trackers is given for Dataset1. In order to compare the beat trackers, a one-way ANOVA followed by a series of t-tests with level of significance of $\alpha = .05$ was performed. Tukey's HSD adjustment was used to account for the effect of multiple comparisons. The most accurate beat tracking results without statistically significant differences are depicted in boldface.

It can be seen from Table I that a subset of beat trackers perform significantly better than most of the others. The set of best beat trackers varies slightly depending on the evaluation measure which is applied. Comparing the individual accuracy values of the approaches with the mean of all beat trackers shown in the last row of Table I we can see that some approaches perform worse than the mean for all evaluation measures. When looking towards finding a subset of committee members we recall the need for accuracy in beat tracking, since poorly performing beat trackers can lead to an over-estimation of difficulty – where all files appear difficult.

Table I: Ground truth performance of each individual BT on Dataset1. Bold numbers indicate best performances.

BT	AMLt (%)	F-measure (%)	Inf. Gain (bits)
Aubio (AUB) [25]	50.6	49.4	1.58
Beatit (BIT) [26]	61.0	52.7	1.62
Beatroot (DIX) [6]	70.8	61.7	1.98
BeatUJaén (BUJ) [27]	41.6	33.9	1.18
Boeck (BOE) [8]	58.7	**66.6**	1.98
Davies (DAV) [5]	**75.9**	**62.8**	**2.25**
Degara (DEG) [9]	**77.7**	**65.3**	**2.26**
Ellis (ELL) [4]	60.0	55.1	1.76
Essentia (ESS) [28]	57.3	51.7	1.43
Hainsworth (HAI) [11]	59.6	51.1	1.84
IBT causal (IB1) [29]	58.0	55.2	1.67
IBT non-causal (IB2) [29]	**73.8**	60.5	1.92
Klapuri (KLA) [10]	**77.7**	**65.5**	**2.32**
Lee (LEE) [30]	26.4	48.8	1.09
Scheirer (SCH) [31]	49.0	56.2	1.69
Stark (STA) [21]	71.0	59.5	2.03
Mean	60.6	56.0	1.79



(a) Information Gain (b) AMLt (c) F-measure

Figure 3: Development of the oracle scores for the three evaluation measures. The performance of the chosen committee is depicted by a cross and the vertical line marks the point with 5 BT in the oracle.

B. Selecting the committee

While in previous work [19] the way we chose the committee members was not documented, we now illustrate the effect of choosing the committee members based on oracle performances as described in Section II-B. The development of the oracle scores are depicted in Figure 3. A saturation effect can be observed when the number of beat trackers in the subset increases, and we decided to limit the number of beat trackers to five (as shown by the vertical dotted line). The order in which algorithms entered the oracle slightly varied between the evaluation measures. We initially decided to choose the five beat trackers based on their average ranking obtained from the three evaluation measures. This gave [KLA, DEG, HAI, BOE, IB2]. This ranking results in a higher diversity of approaches than by ordering according to ground truth performance. For example, the DAV[1] algorithm is not among the best five methods in the oracle. This is caused by similarity between the DAV and DEG algorithms which share the same input feature and tempo detection method. Therefore, once

[1]Note, we use an improved version of the original algorithm [5] which is implemented as a Sonic Visualiser plugin.

DEG has entered the committee DAV offers little additional improvement. However the fundamentally different methods of HAI and the BOE, which are less accurate overall, are able to increase the diversity of the committee.

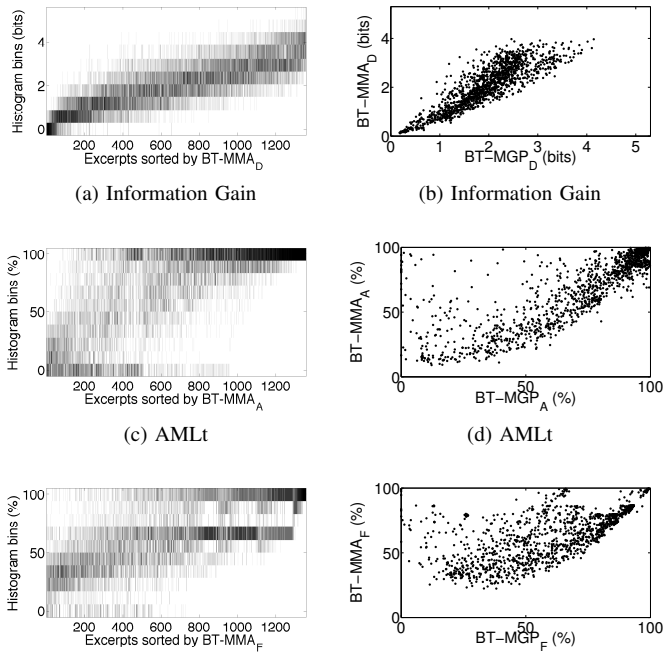
Despite the improvement offered by HAI and BOE, we chose to exclude these approaches from the committee on the grounds of portability, computation time and public availability. Instead, we use the widely available approaches of Dixon (DIX) [6] and Ellis (ELL) [4]. Their inclusion leads to non-significant decrease in oracle performance (marked by a cross in Figure 3) by 0.63%, 0.13% and 1.15% for Information Gain, AMLt, and F-measure, respectively. We hope that the chosen committee: [KLA, DEG, IB2, DIX, ELL] will enable other researchers to most easily reproduce results presented in this paper.

C. MMA computation

After the selection of committee members, mutual agreement between the sequences obtained from the 5 beat trackers were computed using the three evaluation measures described in Section II-A. Then, for each evaluation measure, mutual agreements for a particular piece were summarized in a mutual agreement histogram with 11 bins spanning the whole range of values of the particular evaluation measure (*e.g.* 0% to 100% for AMLt). In the left column of Figure 4 these histograms are depicted for Dataset1. The histograms are sorted by their BT-MMA value for each evaluation method. Dark colors in the histogram plots indicate a high population of the specific histogram bin. In the right column of Figures 4, scatter plots of BT-MMA against mean ground truth performance BT-MGP are shown. For our application, BT-MMA should predict BT-MGP at least for difficult pieces. These are located at low BT-MGP values, while easier pieces are found at higher BT-MGP values, *i.e.* in the region where the beat trackers perform well in the mean for a specific sample.

Comparing the scatter plots for the three evaluation measures we can observe that the BT-MMA$_D$ in Figure 4b is characterized by the highest correlation with the BT-MGP. This correlation is particularly strong for low BT-MMA$_D$ values, which indicates that low BT-MMA$_D$ can reliably predict low ground truth performance. The other two scatter plots (Figures 4d and 4f) show an increased correlation only for high ground truth performance, *i.e.* in the upper right corner of these scatter plots. Based on this evidence it is apparent that F-measure in particular cannot be used to predict poor performance. This differing behavior of Information Gain on the one side and F-measure and AMLt on the other can be attributed to Information Gain having an unambiguous zero value, as shown in Section II-A.

By observing the histogram plots in the left column of Figure 4, it is apparent that only the Information Gain has a continuous transition from histograms centered at low values to histograms centered at high values. The other two measures are characterized by generally flatter histograms, and the F-measure histograms are often characterized by simultaneous high values for 100% and 66.7%. This can be ascribed to beat sequences at metrical levels related by a factor of two (see



(a) Information Gain (b) Information Gain

(c) AMLt (d) AMLt

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

Figure 4: Left side: Each column of the image depicts a histogram obtained from $5 * 4/2$ mutual agreements of the 5 beat sequences for each song in Dataset1. The histograms are sorted by their mean values (BT-MMA). Dark colors indicate high histogram values. Right side: MMA versus MGP scatter plots for each evaluation method.

II-A) which score an F-measure of 66.7%. These characteristics imply that the computation of mean mutual agreement is most reliable for the Information Gain. Hence, we conclude that using Information Gain for the MMA computation is superior to either F-measure or AMLt.

IV. Building a challenging dataset

We start from the assumption that adding diversity to existing collections is necessary to facilitate future improvement in beat tracking systems. To this end we now describe a new dataset and compare its properties to those of Dataset1. The new dataset was compiled by choosing a set of CDs and extracting 40s of each song. We chose music with the goal of obtaining a sufficient number of files that could be considered difficult in terms of their rhythmic properties. We concentrated on styles of Western music, because it is not always apparent how the notion of beat is used in music of other cultures. The CDs contained a variety of styles including classical music, Romantic music, film soundtracks, blues, chanson, and solo guitar compositions. We extracted a total of 678 excerpts.

A subset of the 678 pieces was chosen for manual annotation with the goal of selecting pieces that cause the largest problems to the beat tracking approaches. We decided to choose samples with BT-MMA$_D$ values $\leq$ 1 bit, which resulted in 270 samples. The choice of this threshold was motivated by observing that for values ≤ 1 bit, the histograms in Figure 4a have a clear peak and the correlation with BT-MGP$_D$ in Figure 4b is strong. We do not intend for this

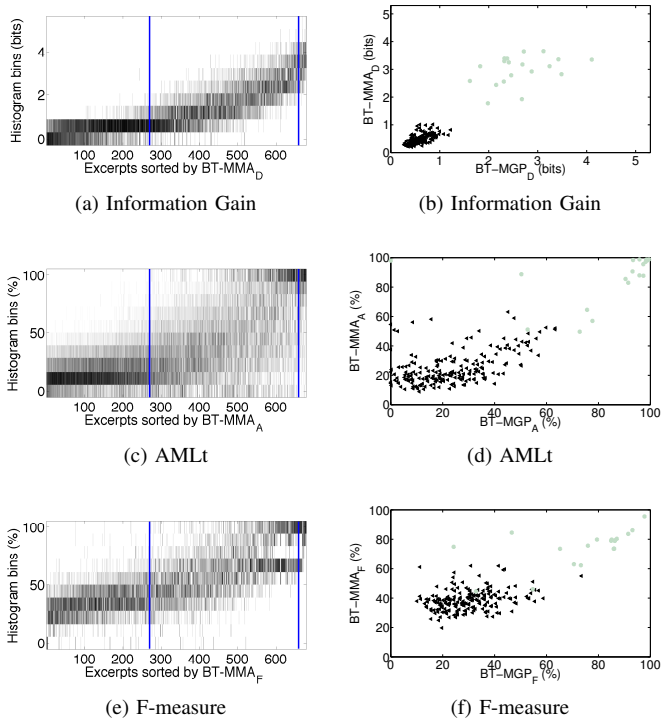
threshold to be interpreted as a globally valid division between easy and difficult files, rather it was chosen empirically to maximize the probability of obtaining only difficult files. In order to cross-check the assumption of these files being difficult, we added 19 samples with the highest BT-MMA$_D$ value which should be characterized by a high BT-MGP. This set of 289 pieces chosen for annotation will be referred to as **Dataset2** throughout the remainder of the paper.

The annotation process followed a detailed protocol, which is available on the paper's website [32]. The first step consists of recording *spontaneous taps* from all authors of this paper for all 289 pieces. The taps enable us to examine the ability of listeners to follow the beat in a possibly difficult piece of music without any entrainment. The MMA of these taps is used to assess the perceptual difficulty, and will be compared to the MMA of the automatic beat trackers. It should be stated that while all five authors come from an engineering background, four have many years experience as practicing musicians in different styles and instruments. Before tapping, each subject was not permitted to listen to the piece, instead they tapped the beat while listening to it for the first time. In addition, no subsequent correction of the taps was allowed.

In the next step, the files in Dataset2 were equally distributed among the authors of the paper for ground truth annotation. The annotations were performed using Sonic Visualiser [33]. To assist with the annotation, each annotator was allowed to use multiple visualizations such as the waveform or spectrogram. The use of automatic beat tracking or onset detection algorithms was not permitted, however the spontaneous taps could be used. Wherever available, scores of the pieces were used as a guideline to arrive at a valid annotation, especially for classical and Romantic music. Each annotator was given the possibility to reject a file if the annotation process appeared intractable. This happened in 72 cases, resulting in 217 valid beat annotations for Dataset2.

Finally, the annotator had to compile a tag file for each annotated sample. The tags specified which signal characteristics made the annotation difficult. An arbitrary number of tags could be assigned to a song, however if the file was not considered difficult for annotation, the tag "none" was used. The full list of tags is presented in Section V-B.

Each annotation was subsequently evaluated by a second subject. In the annotation process all annotators expressed insecurity about some of their annotations due to the high level of difficulty of some of the files. To address this issue we consulted experts with conservatory degrees in music and composition, and with their assistance we obtained a more reliable ground truth especially for the most difficult samples. The comments and changes that were performed in this revision process were documented and are available on the paper's website [32].



(a) Information Gain (b) Information Gain

(c) AMLt (d) AMLt

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

Figure 5: Left side: Each column of the image depicts a histogram obtained from $5*4/2$ mutual agreements of the 5 beat sequences for each song in the 678 samples used to derive Dataset2. The histograms are sorted by their mean values (BT-MMA). Dark colors indicate high histogram values. Files excluded from annotation lie between the vertical blue lines. Right side: MMA versus MGP scatter plots for the annotated 217 files in Dataset2. Pieces assumed to be easy according to their BT-MMA are depicted by gray circles with the remainder shown as black triangles.

V. Analysis of new database

A. Automatic beat tracking on the new dataset

For Dataset2, BT-MMA histograms and scatter plots of BT-MMA over BT-MGP are depicted in Figure 5. Computations were performed in the same way as for Dataset1, enabling a comparison between Figure 4 and Figure 5. A common characteristic of the plots for Dataset1 and Dataset2 is the high correlation between BT-MGP and BT-MMA for small values when using Information Gain (see Figures 4b and 5b), respectively. Again, for F-measure and AMLt such a correlation cannot be observed. This provides strong evidence for using BT-MMA$_D$ to detect difficult files in the context of the newly annotated Dataset2.

Differences between Dataset1 and Dataset2 are evident for all three evaluation measures: the mutual agreement histograms in the left columns are strongly biased towards the upper right corner for Dataset1 and towards the lower left corner for Dataset2. Again, the histograms for BT-MMA$_D$ in Figure 5a show a more accentuated concentration and a continuous development from concentration in low to high histogram bins. However, in Figure 5a a higher proportion of histograms is characterized by a concentration in bins of 1 bit or less. This indicates that Dataset2 contains a larger relative percentage of difficult samples than Dataset1. The super-imposed vertical lines in the histogram plots in Figure 5 indicate the borders for the initial choice of files to be annotated, i.e., the first 270 files and the last 19 files sorted by BT-MMA$_D$ (see Section IV). Samples on the left of the first line were chosen because they were assumed to be difficult

Table II: Ground truth performance of each individual BT on the 217 annotated files in Dataset2. Bold numbers indicate best performances.

BT	AMLt (%)	F-measure (%)	Inf. Gain (bits)
Aubio (AUB)	18.5	24.7	0.68
Beatit (BIT)	20.6	28.7	0.53
Beatroot (DIX)	**27.6**	32.2	0.66
BeatUJaén (BUJ)	23.9	27.7	0.60
Böck (BOE)	**26.1**	**40.1**	**0.91**
Davies (DAV)	**33.4**	32.2	**0.90**
Degara (DEG)	**33.4**	**34.6**	**0.89**
Ellis (ELL)	20.8	**35.2**	0.62
Essentia (ESS)	23.3	26.6	0.64
Hainsworth (HAI)	**26.0**	24.8	**0.83**
IBT causal (IB1)	21.1	26.8	**0.70**
IBT non-causal (IB2)	**28.6**	31.1	**0.78**
Klapuri (KLA)	**33.9**	**36.2**	**0.92**
Lee (LEE)	12.9	**34.6**	0.50
Scheirer (SCH)	18.5	30.2	**0.70**
Stark (STA)	**26.0**	27.3	**0.74**
Mean	22.7	30.8	0.73
Deterministic	16.1	21.2	0.46

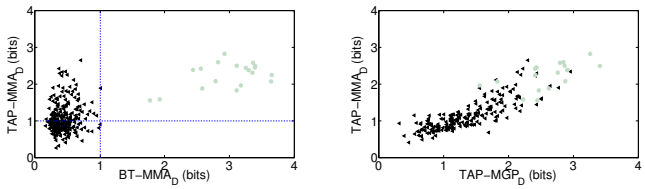
(low BT-MMA$_D$), while the 19 files on the right of the second line in the histogram plots were included because they were supposed to be the easiest in the dataset (high BT-MMA$_D$). In Figure 5b a clear separation can be observed between those files, where the difficult files are marked by black triangles and the easy files by gray circles. This separation is not evident for the other evaluation measures in Figures 5d and 5f, and the difficult files form wider spread clusters.

The individual accuracy values for Dataset2 are depicted in Table II where bold numbers indicate the best beat tracking results without statistically significant differences. Note that the files in Dataset2 were selected based on BT-MMA$_D$ and are supposed to be difficult, with the exception of the included 19 files with high BT-MMA$_D$. For Dataset2 the overall performance is much lower than for Dataset1 (see Table I), and there are fewer significant differences among the best beat trackers. Moreover, there is no consistent subset of best beat trackers, as all except four beat trackers are among the best performers for at least one evaluation method. The performance of some beat trackers is close to the mean performance of an entirely deterministic (baseline) beat sequence, fixed at 120 bpm and generated as in [20]. In general, this proves that the compiled dataset is more difficult for automatic beat tracking than Dataset1, and again supports the validity of our proposed BT-MMA method.

B. Perceptual vs. automatic beat tracking difficulty

1) Assessing perceptual difficulty: To better understand the difficulty of beat tracking, subjective aspects should be taken into account as well. In Dataset2, we can gain insight into these subjective aspects by using the spontaneous taps collected in the annotation process.

During the annotation of Dataset2, we found that spontaneously tapping to an unknown piece is a very demanding process for music without a clear and simple beat. Thus, we assume that perceptually easier files result in tap sequences that show higher mutual agreement, analogous to the beat tracker outputs. In order to differentiate these agreements from the MMA obtained from beat trackers (i.e. BT-MMA) we will refer to them as **TAP-MMA**, and to the mean performance of the taps compared to ground truth as **TAP-MGP** (in contrast to BT-MGP). The TAP-MMA values between the five spontaneous taps that are available for each sample were computed using Information Gain. Figure 6a shows a scatter plot of these TAP-MMA$_D$ values against the BT-MMA$_D$ values of the five beat tracking algorithms. While the sparse cluster in the upper right corner indicates that high agreement of beat sequences implies high agreement of spontaneous taps, such a relation does not exist for low BT-MMA$_D$. In this case, we can observe the existence of a wide range of TAP-MMA$_D$ values. This implies that among files that are difficult for automatic beat tracking, there were both difficult and easy files for the human tappers. In Figure 6b a high correlation between TAP-MMA$_D$ and the mean performance of the taps against the ground truth annotations (TAP-MGP$_D$) can be observed. This correlation supports the assumption that high agreement between subjects implies perceptually easier pieces. Comparing Figures 5b and 6b, we can see that in Figure 6b there are no separate clusters of data for very low TAP-MMA$_D$ and TAP-MGP$_D$values. This indicates that, for the difficult samples, the human taps tended to be more accurate compared to the ground truth, and that the spontaneous taps were characterized by higher mutual agreement than the beat tracker outputs.



(a) (b)

Figure 6: TAP-MMA$_D$ and TAP-MGP$_D$ for annotated 217 files in Dataset2. Pieces which are considered easy according to their BT-MMA$_D$ are depicted by gray circles. (a): Scatter plot of TAP-MMA$_D$ versus BT-MMA$_D$, dotted lines indicate the chosen border for difficult files for beat tracking (vertical line) and human tappers (horizontal line) (b): Scatter plot of TAP-MMA$_D$ versus TAP-MGP$_D$

In conclusion, we can state that, even without ground truth available, it is possible to reliably detect samples where automatic beat tracking will fail. Among these files there will be both files that are perceptually difficult and files that are easy. As our aim is to facilitate improvement in beat tracking, we want to focus on those pieces that have a perceivable beat but that make beat trackers fail. These pieces are located in the top-left rectangle of Figure 6a, and we will now focus on the signal properties that differentiate them from perceptually difficult pieces which are located in the lower-left rectangle of Figure 6a.

2) Signal properties: The general signal properties encountered in Dataset2 are summarized in the tags assigned

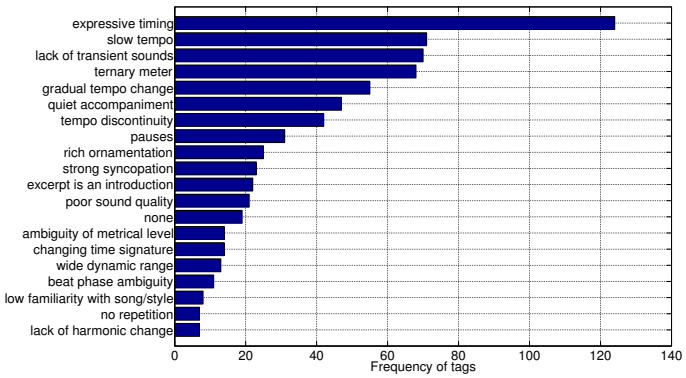


Figure 7: Frequency of tags for all annotated files in Dataset2. Tags indicate which signal properties made a sample appear difficult during the manual annotation.

during the annotation process. Figure 7 shows the number of occurrences of all tags for the 217 annotated pieces. The most prominent tag is *expressive timing*, which applies when a sample changes in tempo in correlation with its melodic phrase or segment boundaries [34] as often happens in Romantic music. Other prominent tags related to tempo were *slow tempo*, *gradual tempo change* (*i.e.* one stable tempo changes gradually to a different stable tempo) and *tempo discontinuity* (*i.e.* a sudden tempo change). This indicates that any kind of tempo changes cause trouble for beat tracking approaches, and adds the characteristic of having a slow tempo to the list of problematic tempo-related features. Furthermore, ternary meter also led the beat trackers to fail, which suggests that many approaches may be tailored to track music mainly in a $\frac{4}{4}$ time signature. Characteristics related to the instrumental timbres, such as *lack of transient sounds* and *quiet accompaniment* complete the picture of the problematic signal properties that make beat trackers fail. They can be summarized in three groups: i) timing/tempo related, ii) time signatures and iii) lack of clear rhythmic onsets. The tag *none* was applied when none of the other tags fit to the properties of the signal, and its appearance is always related to the files with high BT-MMA$_D$, *i.e.* the 19 easy files in Dataset2.

Having obtained an overview of the signal properties that make automatic beat tracking difficult, we would like to know which of these properties makes tapping the beat difficult for human listeners. We want to address the question of whether the files in the upper and lower left rectangles of Figure 6a differ according to their signal properties. If we can identify some significant differences, this can give valuable insight into how to discriminate between perceptually difficult pieces and those that are difficult only for automatic beat tracking. To this end, features describing those discriminant signal properties might be used in a machine learning approach to automatically classify samples into one of the two classes. A threshold was set to a TAP-MMA$_D$ value of 1 bit (dotted horizontal line in Figure 6a), *i.e.* the same threshold that was applied to BT-MMA$_D$ when choosing difficult files for annotation. Then, a set of t-tests was applied in order to investigate if the beat-annotated samples in the lower and upper left rectangles differed regarding their given tags. In this way, we can infer which signal properties led to inaccurate tappings.

Table III: Tags with different mean according to t-test, sorted by increasing p-value, from top to bottom. The presence of a tag implies that it appears significantly more frequently for low TAP-MMA$_D$

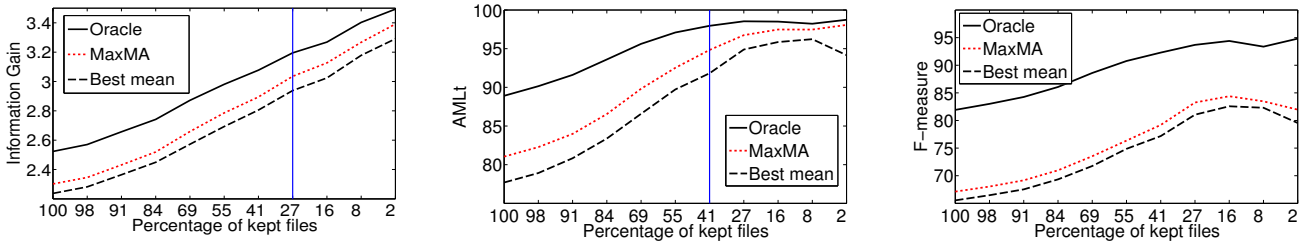
T-test: TAP-MMA$_D$	p-value
changing time signature	0.0010
expressive timing	0.0011
quiet accompaniment	0.0035
no repetition	0.0047
low familiarity with song/style	0.0110
beat phase ambiguity	0.0360

The results of the t-tests are listed in Table III. The appearance of a tag in the list means that it is significantly more present in files with low TAP-MMA$_D$. We can see that a change in time signature was the most important factor that led to low tapping agreement. However, this tag is quite sparse among the dataset as shown in Figure 7. The most prominent factors, taking into account their frequency of appearance, are expressive timing and quiet accompaniment. Hence, these factors apparently cause problems both for beat trackers and for human tappers. The list of properties given in Table III can serve as a guideline to which signal descriptors might be applied when trying to exclude signals from automatic beat tracking because of their high complexity even for human listeners. It is apparent that processing music with highly expressive timing should be postponed, as its beat is too complex to be spontaneously tracked even by human listeners. We consider that demanding an accurate beat tracking on such music resembles demanding high word recognition rates from an automatic speech recognizer in signals that cannot be perceived by a human listener. However, a profitable first step may be to concentrate on music characterized by ternary meters, slow tempo or soft onsets, among other characteristics that do not impose drastically increased difficulty to human beat perception.

VI. Sample application

In this section, we demonstrate a sample application for the mutual agreement technique that is different from sample selection in compiling datasets. We assume a large collection of audio files without any beat annotations and we would like to perform a task that relies on beat tracking, e.g. cover song detection or a chord transcription. As a first step, we seek to reject any files considered impossible for current beat tracking systems. Then for the remainder, we would like to choose a reliable beat tracker to provide the beats. Traditionally this would be done by selecting an algorithm which is considered superior to the others based on some beat tracking evaluation process. We now show how our mutual agreement measure with five beat tracking algorithms can be applied for this purpose as well.

In this experiment we ran our committee of beat trackers on Dataset1 and calculated BT-MMA$_D$ for each sample. We then excluded those samples with BT-MMA$_D$ below a specified



(a) Information Gain (b) AMLt (c) F-measure

Figure 8: Result of automatic beat tracker selection (MaxMA), compared with single best beat tracker choice (Best mean) and oracle scores (Oracle) on Dataset1 using our committee of 5 BT. For the thresholds 0 to 3 bits on BT-MMA$_D$, the percentage of the 1360 files kept for evaluation is shown on the x-axis. The vertical line shows the point up to which differences between MaxMa and Beat mean are significant.

threshold. The threshold was incremented in steps of 0.3 bits from 0 to 3 bits. Since we have shown that disagreement between the committee harms beat tracking performance, we now try the opposite approach and select the beat sequence with maximum mutual agreement with the committee, which we denote, **MaxMA**. For each sample (at a given threshold), we simply select the beat sequence with the maximum mutual agreement (MaxMA) with the other four sequences as the most reliable beat estimation. In effect we assume that the beat tracker that best agrees with the rest of the committee is the most reliable algorithm. In Figure 8, we compare the MaxMA approach to another viable option, that of picking the beat tracker [10] with the best mean overall performance from our experiments in Sections III and V. We denote this option **Best mean**. To illustrate the upper limit on performance we also include the theoretical optimum **Oracle**, that picks the most accurate beat tracker for each individual sample.

Figure 8 shows that applying the MaxMA method to choose a beat tracker leads to significant improvements when evaluated against ground truth for both Information Gain and AMLt over a wide range of thresholds. T-tests with a level of significance of $\alpha = .05$ were performed to compare the MaxMA with the Best mean at each threshold, and all differences on the left of the vertical lines in Figures 8a and 8b are significant. This improvement in performance occurs even when no samples are discarded and remains when retaining up to 41% for of samples AMLt and 27% for Information Gain. Beyond this point only the samples with high mutual agreement remain, which are among the easiest in the dataset, hence the choosing MaxMA over the the Best mean may offer less improvement. Indeed both the MaxMa and Best Mean performance approach the Oracle when only very few (easy) samples remain.

While there is still a consistent improvement for the F-measure (Figure 8c), this improvement is not significant for any threshold value. This is likely the result of the discontinuity of the F-measure, which assigns 0% to beat sequences misaligned in phase and values of 66% for tempo halving/doubling. These properties of the F-measure increase its variance even for sets of beat sequences that can be acceptable in terms of perceptual criteria. This supports the observation that significant differences in beat tracking performance can vary dependent on the evaluation measure [20].

On the basis of this sample application, we infer that mutual agreement can be successfully applied both for choosing "beat-trackable" files and for improving beat tracking performance on these files by selecting the beat tracker that has the maximum mutual agreement with the other beat trackers. Since all beat sequences must be estimated for the file selection/rejection process, the improvement given by the MaxMA beat tracker choice adds negligible additional complexity.

VII. Conclusions

In this paper, we presented a method based on mutual agreement of beat sequences to detect informative samples in non-annotated data collections. We compiled and annotated a new dataset that consists mainly of pieces with low mutual agreement, and showed that this dataset is significantly more difficult for state of the art beat tracking algorithms than the largest existing collection. Using the new difficult dataset, we analyzed the signal characteristics that make beat trackers fail, and investigated the extent to which these characteristics coincide with the properties that make tapping difficult for humans. Based on our informal analysis of human tapping it appears that expressive timing contributes strongly to making music difficult to tap to. Furthermore it may not be musically appropriate to attempt to precisely follow large expressive changes. The musical experts who assisted in the annotation process demonstrated more musically meaningful annotations could be obtained by tapping a stable pulse around which the timing changes deviate. However this level of tapping required extensive musical training (beyond the level of the authors) and provides strong evidence towards rejecting beat tracking for musical pieces of this nature. Towards more realistic advances in beat tracking, we propose investigating techniques for music with properties that do not pose such considerable difficulties for humans, including pieces characterized by ternary meter, slow tempo, or soft instrument onsets.

In order to reliably detect difficult samples using mutual (dis-)agreement, we demonstrated that the choice of the evaluation measure is crucial, and that Information Gain was better suited to this task than both the F-measure and AMLt evaluation methods. However, Information Gain appears less effective in highlighting where beat tracking algorithms strongly agree with each other. Hence, in future work, we will explore methods to combine different evaluation methods.

The proposed MMA method represents an efficient approach to improve diversity in existing datasets, as well as a simple technique to improve beat tracking in large non-annotated datasets. Our method can also be applied in other contexts by detecting problematic files for chord recognition where it may be valuable to reject the use of beat tracking as a temporal analysis component. Furthermore, outside of beat tracking, we believe that there is considerable scope to apply mutual agreement to other MIR research tasks through the use of context specific evaluation methods.

The audio files of the newly compiled beat tracking dataset will be made available on request, and all of the accompanying meta-data is available on the paper's web-page [32]. We encourage the research community to contribute to this resource by adding further annotated difficult samples along with meta-data.

VIII. Acknowledgements

We thank the authors of the beat tracking algorithms for making their code available. For assistance in improving the annotations, we thank Michael Hecht and the group at Butler School of Music in UT Austin. We would also like to thank Jeremy Pickens for inspiring discussions.

References

[1] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[2] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 2009, pp. 231–236.

[3] A. Holzapfel and Y. Stylianou, "Parataxis: Morphological similarity in traditional music," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 2010, pp. 453–458.

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

[5] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020, 2007.

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

[7] G. Peeters, "Beat-tracking using a probabilistic framework and linear discriminant analysis," in *Proceedings of the 12th International Conference on Digital Audio Effect, DAFx-09*, 2009, pp. 313–320.

[8] S. Böck and M. Schedl, "Enhanced Beat Tracking with Context-Aware Neural Networks," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011, pp. 135–139.

[9] N. Degara, E. Argones, A. Pena, S. Torres-Guijarro, M. E. P. Davies, and M. D. Plumbley, "Reliability-informed beat tracking of musical signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 290–301, 2012.

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

[11] S. W. Hainsworth and M. D. Macleod, "Particle filtering applied to musical tempo tracking," *Journal of Advances in Signal Processing*, vol. 15, pp. 2385–2395, 2004.

[12] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[13] G. Tzanetakis and G. Essl, "Human perception and computer extraction of musical beat strength," in *Proceedings of the 12th International Conference on Digital Audio Effect, DAFx-02*, 2002, pp. 257–261.

[14] P. Grosche, M. Muller, and C. S. Sapp, "What Makes Beat Tracking Difficult? A Case Study on Chopin Mazurkas," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 2010, pp. 649–654.

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

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

[17] 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.

[18] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, "Support vector machine active learning for music retrieval," *Multimedia systems*, vol. 12, no. 1, pp. 1–11, August 2006. [Online]. Available: http://mr-pc.org/work/mmsj05.pdf

[19] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, "On the automatic identification of difficult examples for beat tracking: towards building new evaluation datasets," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, 2012, pp. 89–92.

[20] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Queen Mary University of London, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06, 2009. [Online]. Available: http://www.elec.qmul.ac.uk/people/markp/2009/DaviesDegaraPlumbley09-evaluation-tr.pdf

[21] A. M. Stark, M. E. P. Davies, and M. D. Plumbley, "Real-time beat-synchronous analysis of musical audio," in *Proceedings of the 12th International Conference on Digital Audio Effect, DAFx-09*, 2009, pp. 299–304.

[22] J. P. Bello, "Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats," in *Proceedings of the 8th International Conference on Music Information Retrieval*, 2007, pp. 239–244.

[23] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Measuring the performance of beat tracking algorithms using a beat error histogram," *IEEE Signal Processing Letters*, vol. 18, no. 3, pp. 157–160, 2011.

[24] F. Gouyon, "A computational approach to rhythm description — Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing," Ph.D. dissertation, Music Technology Group, Universitat Pompeu Fabra, 2005. [Online]. Available: http://www.iua.upf.es/ fgouyon/thesis/

[25] P. M. Brossier, "Automatic annotation of musical audio for interactive systems," Ph.D. dissertation, Department of Electronic Engineering, Queen Mary University of London, 2006. [Online]. Available: http://aubio.org/phd/

[26] J. Bonada and F. Gouyon, "Beatit," 2006, mtg.upf.edu, internal software.

[27] R. Mata-Campos, F. J. Rodriguez-Serrano, P. Vera-Candeas, J. J. Carabias-Orti, and F. J. Canadas-Quesada, "Beat tracking improved by am sinusoidal modeled onsets (1) - mirex 2010," in *Music Information Retrieval Evaluation eXchange (MIREX)*, no. 1, 2010.

[28] E. Aylon and N. Wack, "Beat detection using plp," in *Music Information Retrieval Evaluation eXchange (MIREX)*, 2010.

[29] J. L. Oliveira, F. Gouyon, L. G. Martins, and L. P. Reis, "IBT: A real-time tempo and beat tracking system," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 2010, pp. 291–296.

[30] T. C. Lee, "MIREX 2010 Audio Beat Tracking Program," in *Music Information Retrieval Evaluation eXchange (MIREX)*, 2010.

[31] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[32] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. Oliveira, and F. Gouyon, "Webpage for selective sampling for beat tracking evaluation," February 2012. [Online]. Available: http://smc.inescporto.pt/research/data

[33] C. Cannam, C. Landone, and M. Sandler, "Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia International Conference*, 2010, pp. 1467–1468.

[34] N. Todd, "A computational model of rubato," *Contemporary Music Review*, vol. 3, no. 1, pp. 69–88, 1989.