ASSIGNING A CONFIDENCE THRESHOLD ON AUTOMATIC BEAT ANNOTATION IN LARGE DATASETS

First author
Affiliation1
author1@ismir.edu

Second author
Retain these fake authors in
submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

In this paper we establish a threshold for perceptually acceptable beat tracking based on the mutual agreement of a committee of beat trackers. In the first step we use an existing annotated dataset to show that mutual agreement can be used to select one committee member as the most reliable beat tracker for a song. Then we conduct a listening test using a subset of the Million Song Dataset to establish the threshold which results in acceptable quality of the chosen beat output. For both datasets, we obtain a percentage of trackable music of about 73%, and we investigate which data tags are related to acceptable and problematic beat tracking. The results indicate that current datasets are biased towards genres which tend to be easy for beat tracking. The proposed methods provide a means to automatically obtain a confidence value for beat tracking in non-annotated data and to choose between a number of beat tracking outputs.

1. INTRODUCTION

Beat tracking can be considered one of the fundamental problems in music information retrieval research. There have been numerous algorithms presented (e.g. [5,6,8,11]) whose common aim is to "tap along" with musical signals. Furthermore the inclusion of beat trackers within other music analysis tasks (such as harmony analysis [9], structural segmentation [12]) has become common-place. However despite the somewhat "automatic" inclusion of beat trackers as temporal processing components, beat tracking itself is not considered a solved problem. Recent comparative studies of beat trackers suggest there is often little to choose between the best performing state of the art methods [4, 13]. Indeed the viewpoint could be taken that beat tracking performance is approaching a glass ceiling of sorts [10] with the accuracy of current algorithms stagnating at around the 80% accuracy mark when evaluated using the least stringent evaluation metrics in common datasets [4].

In previous work [10] we proposed that the presence of this apparent glass ceiling was not confined to beat tracking algorithms having reached their full potential, but also that the datasets on which beat trackers are evaluated simply do not contain a sufficient proportion of challenging examples; and that current beat trackers have over-learned the musical properties of the "easier" excerpts within these datasets. Towards the future advancement of beat tracking we presented a technique to automatically identify challenging examples for beat tracking without the need for ground truth annotations [10]. Our technique was based on measuring the mean mutual agreement (MMA) between a committee of state of the art beat tracking algorithms, where by low mutual agreement (or put another way, high disagreement) between beat outputs was shown to be a good indicator of low accuracy against the ground truth. To this end we determined an MMA "failure" threshold below which beat tracking performance was shown to be very poor, and created a new database comprised entirely of challenging excerpts with MMA below this threshold.

In this paper we address the opposite issue, instead of trying to find where beat tracking algorithms fail, we wish to identify when beat tracking has been successful. When ground truth annotations are available this question can be easily answered, however the problem is non-trivial when no ground truth exists, *i.e.*, on the vast majority of music. The current implicit means for doing so is simply to extrapolate the performance on the limited dataset for which a precise evaluation can be conducted, and assume this is representative of the performance of beat tracking on all music.

In light of our previous concerns about the make up of these annotated databases, we believe that extrapolating performance in this way will give an optimistic estimate of performance. Therefore when seeking to determine an unbiased measure of performance we can either annotate more and more music examples for evaluation, or instead attempt to estimate beat tracking performance without ground truth. If no ground truth is required, then performance can be estimated on very large (effectively unlimited) collections of music. Due to the impractical nature of the first option, we pursue the second.

We extend our previous work to determine an MMA "success" threshold above which we can have high confidence in the beat tracking output of a committee of state of the art beat trackers. We determine the success threshold by means of a subjective listening test, where listeners are asked to rate the quality of the beat output given by the

committee across a range of excerpts for which the MMA has been calculated. In each case the beat tracking output chosen to represent the committee is selected automatically as the beat sequence which most agrees with the remainder of the committee, *i.e.*, the sequence with the MaxMA (see Section2). We illustrate that selecting between beat tracking sequences using MaxMA leads to improved performance over picking any individual algorithm from the committee.

Through the calculation of both MMA and MaxMA we present a technique by which we can estimate the level of successful beat tracking on any dataset without ground truth, and for those excerpts with MMA above the threshold, automatically annotate the beats in a way that exceeds the performance of the state of the art. In light of the recently presented Million Song Dataset [1] we consider this work to be particularly timely.

The remainder of the paper is structured as follows; Section 2 gives an overview of up the proposed method based on mutual agreement and describes the chosen committee and the way to compute the mutual agreement between the members. Section 3 demonstrates the improvement when selecting a beat tracker based on the MaxMA approach on an annotated dataset. Section 4 applies the technique to non-annotated data and describes the procedure followed in the listening test as well as the results of the listening test along with the tag analysis. Section 5 concludes the paper with discussion and future work.

2. MEASURING MUTUAL AGREEMENT

The measure of Mean Mutual Agreement MMA is inspired by the Query by Committee concept [15] which selects the most informative set of samples from a database based on the mutual (dis-)agreement between a designated committee of learners. As depicted in Figure 1, the MMA of a sample is computed by using the beat estimations of a committee of N beat trackers on a musical piece, measuring the mutual agreement $MA_{i,j}$ (see eq. (4)) between estimated beat sequences i and j, and retrieving the mean of all $N(N-1)/2$ mutual agreements:

$$\text{MMA} = \frac{1}{N(N-1)/2} * \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} MA_{i,j}. \quad (1)$$

In addition to calculating the MMA as a summary statistic, we can easily identify the mutual agreement, MA_i, of the beat sequence i which mostly agrees with the remainder of the committee, the MaxMA, and the beat sequence i which agrees the least, the MinMA:

$$MA_i = \frac{1}{N-1} * \sum_{j=1, j\neq i}^{N} MA_{i,j}, \quad (2)$$

$$\begin{cases} MaxMA = max_i\,(MA_i) \\ MinMA = min_i\,(MA_i) \end{cases}, \quad (3)$$

where $i, j = 1, \ldots, N$ and $i \neq j$. In order to measure the mutual agreement $MA_{i,j}$ between each pair $\{i, j\}$ of estimated beat sequences a beat tracking evaluation criteria must be selected. In [10] we reviewed the properties of existing evaluation measures in the literature [2] and selected the Information Gain approach [3] (InfGain) as the only measure with a true zero value, able to match low MMA (measured in bits) with unrelated beat sequences:

$$MA_{i,j} = InfGain(i, j), \qquad i, j = 1, \ldots, N \wedge i \neq j. \quad (4)$$

We selected a committee of five state of the art and publicly available beat trackers: Dixon (Dix.) [5], Degara (Deg.) [4], Ellis (Ell.) [6], IBT [14], and Klapuri (Kla.) [11]. These convey the accuracy and diversity necessary to compute a reliable MMA [10].

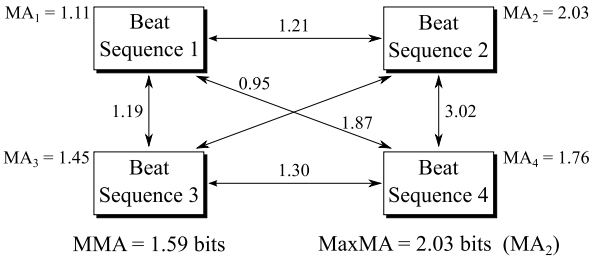


Figure 1: Example calculation of the MMA and MaxMA for a musical piece with the beat sequences estimated from a committee of four beat trackers.

3. MUTUAL AGREEMENT ON EXISTING ANNOTATED DATA

In order to assess if the mutual agreement among our committee of beat trackers can reliably inform us about the best estimated beat sequences we computed and compared the outputs of this committee on a beat annotated dataset containing 1360 excerpts [5, 7] (referred to as **Dataset1360**). Our hypothesis is that the beat tracker that best agrees with the rest of the committee will be the most reliable algorithm for a specific musical piece. Therefore, we compare the mean groundtruth performance of the best beat tracker (Best Mean) against the mean scores of the algorithms responsible for the MaxMA and MinMA for each musical piece. In addition, we compute the Oracle as the mean score given by the best beat sequence estimated by our committee for each musical piece, *i.e.*, the upper limit on performance.

Figure 2 compares the results of the described performance variants on Dataset1360. As described in Section 2, the MaxMA and MinMA were measured using the InfGain [1]. In order to compare these measures against the Best Mean and Oracle performances of the committee on the same data we relied on the AMLt [1] (Allowed Metrical Level with no continuity required) score, as described in [3, 8]. This provides a measure of performance on a more intuitive scale of 0 to 100% and allows some ambiguity in the choice of metrical level at which the beats are tapped.

[1] the InfGain and AMLt measures were computed using the the beat tracking evaluation toolbox, available at `http://code.soundsoftware.ac.uk/projects/beat-evaluation`

The results were computed for different amounts of data confined by incremental values of MMA, in the range of [0-3] bits and varying in steps of 0.3 bits. These MMA values act as a threshold for the selection of music files from the dataset (*e.g.*, for an MMA of 2.1 bits we retain 52.1% of the song excerpts in the dataset).

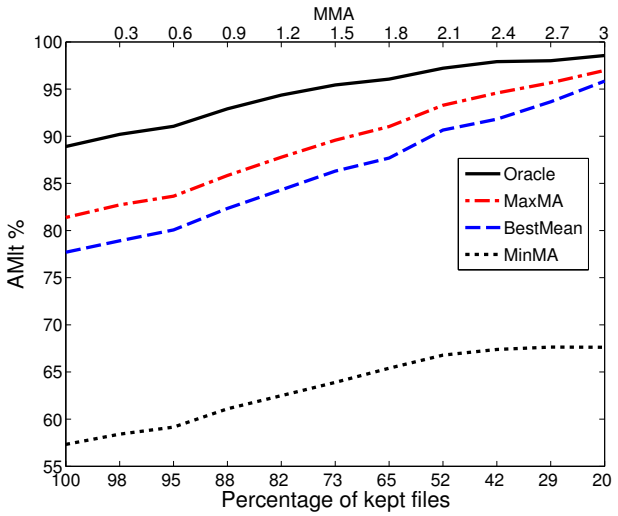


Figure 2: AMLt scores of the beat sequence with maximum (MaxMA) and minimum (MinMA) agreement per song, compared with the single best beat tracker choice (BestMean), and the oracle score (Oracle) for various thresholds of MMA applied to the Dataset1360.

Figure 2 depicts the AMLt scores for different objective choices of beat sequences estimated from our committee of beat trackers on Dataset1360, over decreasing amount of data confined by increasing MMA thresholds. As expected, the overall performance of the committee increases with the MMA threshold. This confirms the hypothesis that the MMA is able to reliably detect difficult files for beat tracking, and therefore can confine the data to easier files by removing those with low MMA. Across all MMA thresholds it is clear that, although lower than the Oracle, the performance of the MaxMA beat sequence outperforms the BestMean algorithm. The difference between the MaxMA and BestMean, of around 3.3%, is statistically significantly ($p<0.05$) for all the excerpts with an MMA below 2.4 bits. Above 2.4 bits this difference is no longer significant and all the variants' performance tendentiously very high. This suggests that below some difficulty extent an objective choice between members of the committee no longer provides any additional advantage in terms of beat tracking accuracy. It is also clear that this performance is highly degraded when picking the MinMA instead.

In conclusion, these results confirm the hypothesis that a good choice of beat trackers is intrinsically dependent on the level of agreement of their estimated beat sequences with the others.

4. AUTOMATICALLY BEAT-ANNOTATING A LARGE DATASET

Having illustrated the validity of using the MaxMA method to select a beat tracking output among a committee of algorithms on an annotated dataset, we now turn our attention to applying it to a large collection of non-annotated data. For very large collections it is impractical to expect there to be ground truth annotations on which to base the performance evaluation. Towards understanding how well the state of the art in beat tracking can annotate beats in large collections we employ our MMA and MaxMA methods and attempt to determine the proportion of excerpts for which the beat estimates are acceptable via a subjective listening test. We want to establish a threshold on MMA above which the beat tracking outputs are perceptually acceptable. For each file, the beat tracking output will be chosen using the MaxMA method.

4.1 Million Song Subset

The large collection we aim to automatically annotate is the **MillionSongSubset** from the Million Song Dataset [1]. The subset is comprised of 10,000 files for which audio previews were obtained. The majority of audio previews were either 30 s or 60 s in duration. To provide sufficiently long excerpts for beat tracking we discarded any audio excerpts shorter than 20 s. This left a set of 9940 excerpts on which to annotate beats. To complement the audio data, we obtained 31696 Last.fm [2] tags covering a subset of 4638 songs.

Once all of the audio and meta data was collected we ran the committee of beat tracking algorithms recording the MMA value per excerpt and saving the MaxMA beat sequence.

4.2 Subjective Listening Test

The aim of our listening test was to determine an MMA threshold above which the beat sequence given by the MaxMA method was deemed acceptable to human listeners. By subsequent inspection of the number of files in the dataset above the MMA threshold we could estimate the proportion for which beat tracking can be considered successful.

Just as it is not possible to hand annotate beats in nearly 10,000 audio excerpts, it is equally impractical to ask participants to listen and rate this large number of excerpts. As alternative to an exhaustive rating of all audio excerpts we selected 8 levels of MMA = [0.5, 1.0, 1.5, . . . , 4.0] bits and chose 6 excerpts per MMA level, giving a total of 48 excerpts to summarize the dataset. To create the musical stimuli for the listening test we constructed stereo audio files containing a mixture of source audio and the MaxMA beat output synthesized as short click sounds. To mitigate the effect of errors in beat tracking at the start of

[2] http://labrosa.ee.columbia.edu/millionsong/lastfm

excerpts, which might bias the listener ratings, each musical stimuli was formed out of the middle 15 s of each excerpt. To allow listeners to hear the audio with and without click sounds, we panned the source audio on its own on the left channel, and on the right channel we mixed the click sounds conveying the beats with a quiet version of the source audio. Through informal listening tests prior to the main experiment, this was deemed an acceptable method for creating the stimuli.

To take the listening test we recruited 25 participants (21 male, 4 female) with an age range of 23 to 41 (mean = 31 years, std = 4.7 years). The participants' level of music training ranged from 0 to 20 years (mean = 8.7 years, std = 7.7 years). Each participant was instructed to perform the test in a quiet environment with good quality headphones. Prior to starting the main test, the participants were given three training examples (not in the main set of 48). The training phase was used for three reasons: *i)* to familiarise participants with the type of musical stimuli in the test, *ii)* for the participants to understand the panning of the beats in the stimuli and *iii)* so the participants could set the playback volume to a comfortable level. To prevent order effects in the stimuli, each participant was given a playlist of files in a different random order.

In taking the test, the participants were asked to answer the following question: "*How do you rate the overall quality of the given click as a beat annotation of the piece?*" The options for rating are: 1 - Bad, 2 - Poor, 3 - Fair, 4 - Good, 5 - Excellent.

4.3 Results

Figure 3 presents a comparison between the human ratings and the MMA of our committee of beat trackers for the selected 48 pieces of the MillionSongSubset. The plot shows that for an MMA equal to 1.5 bits the mean ratings are of 3.7 (Good) with a standard deviation of 0.93. The choice of this value as our threshold of perceptual confidence is confirmed by the highly significant difference ($p < 0.0001$) between the mean ratings achieved for an MMA=1 bits, of 2.4 (Poor), and for an MMA=1.5 bits, of 3.7 (Good).

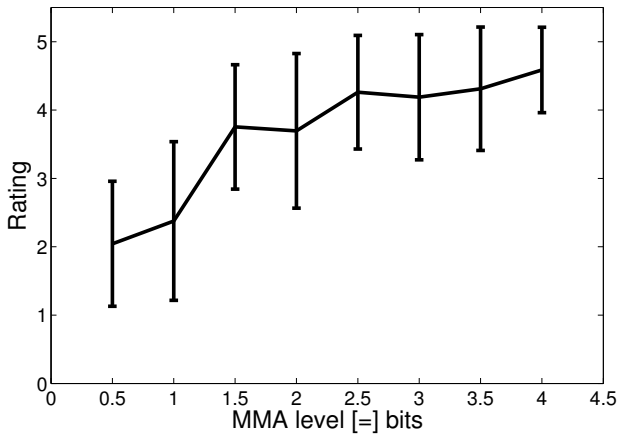


Figure 3: Listening test ratings *vs* MMA for the selected 48 music excerpts, with 15 s each, from the MillionSongSubset.

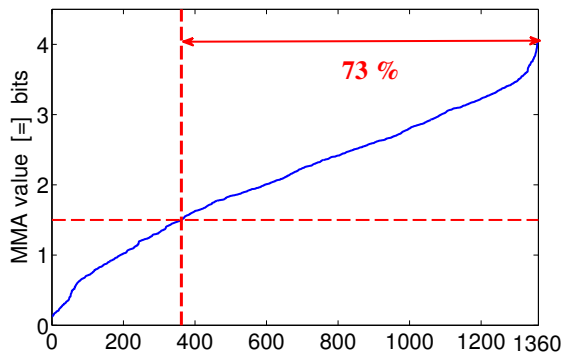
By selecting an MMA of 1.5 bits as a threshold of perceptual confidence for reliable beat tracking we find 996 songs (73%) in Dataset1360 and 7252 audio files (73%) in MillionSongSubset above this limit (see Figure 4). Table 1 presents the AMLt score of the Oracle, the MaxMA, the Best Mean, and the MinMA for the two subsets of Dataset1360 separated by MMA = 1.5 bits, against their groundtruth. The beat tracking performance is consistently high for files with MMA >1.5 bits, with a mean MaxMA performance of ≈90%, which must be considered very accurate, and hence hints at a meaningful relationship between subjective judgement of beat tracking and the AMLt scores obtained from objective evaluation. While beat tracking performance is lower for MMA $<$ 1.5 bits this doesn't mean the MaxMA beat estimations cannot be accurate, but rather that we do not have high confidence in them.

Figure 5 presents histograms for both evaluated datasets depicting the proportion of files where each algorithm is selected as the MaxMA beat sequence. Both histograms show similar shapes, indicating there may be some similar properties between the musical content of both datasets. The two most chosen algorithms are those of Degara and Klapuri; both of which perform most accurately against the ground, and perhaps can be considered at the top end of the state of the art methods. As to why the Degara algorithm is chosen more frequently than that of Klapuri, we can observe in [4] that the inter-quartile range of the Degara algorithm is tighter than that of Klapuri (for a similar median), implying it is "wrong" in a lower proportion of excerpts.

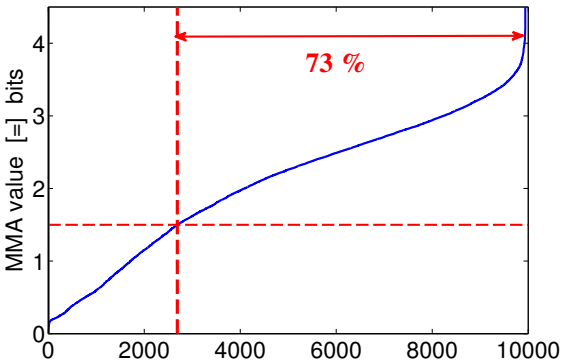
Name	AMLt (%)	MMA
Oracle	95.4	MMA>1.5
MaxMA	89.9	
Best Mean	86.3	
MinMA	63.9	
Oracle	70.9	MMA<1.5
MaxMA	58.8	
Best Mean	54	
MinMA	50.1	

Table 1: Mean AMLt score of Oracle, MaxMA, Best_Mean, and MinMA for the two subsets of Dataset1360 divided by an MMA threshold of 1.5 bits.

Given the MMA threshold, we can now look at the genre-related tags of the files that appear significantly more often (with $p < 0.0001$) in the MillionSongSubset with MMA above and below 1.5 bits. These are shown in Table 2. From inspection of the table we can see that the genres above the MMA threshold are those which we would typically associate with being "easier" for beat tracking where as those below the threshold appear more challenging. Seeing genre labels related to metal music was a surprising result, since this music is strongly percussive and isn't characterised by wide tempo changes. The fact that metal music consistently falls below the threshold indicates it might be the "noisy" element of the music which causes it to be difficult. To the best of our knowledge there are very few



(a) Dataset 1360



(b) MillionSongSubset

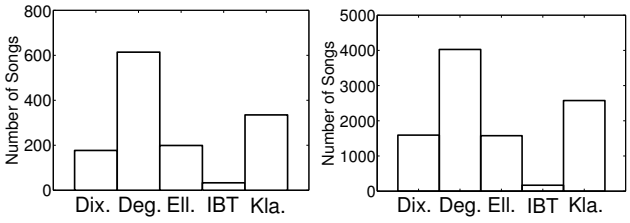
Figure 4: Datasets sorted by MMA and the perceptual threshold of 1.5 bits.

Tag	Frequency	MMA
Rock	1080	MMA>1.5
Pop	680	
Dance	320	
Hip-hop	271	
Rap	193	
Pop rock	154	
Reggae	149	
Jazz	227	MMA<1.5
Instrumental	199	
Death metal	80	
Black metal	74	
Progressive metal	59	
Classical	36	
Grindcore	28	

Table 2: Frequency of the genre-based occurrence of tags for the two subsets of MillionSongSubset divided by an MMA threshold of 1.5 bits.

metal examples in existing beat tracking databases, suggesting it is something of a forgotten genre for beat tracking.

Another important observation relates to the tag frequency for genre labels above and below the threshold. There is a far higher proportion of songs tagged "Rock" and "Pop" compared to all the others, and in general the



(a) Dataset 1360 (b) MillionSongSubset

Figure 5: Histograms with the number of times each algorithm is chosen with the MaxMA approach.

tags used above the threshold appear much more frequently than those below it. From this we can infer that, just as Dataset1360 is biased towards easier cases for beat tracking [10], the same could be said of the MillionSongSubset. Evidence for this conclusion can be found in the description of the MillionSongDataset itself [1] where the lack of diversity is mentioned - in particular the small amount of classical and world music.

Given the disproportionate number of easier beat tracking files in this dataset, our estimate of 73% of files for which beat tracking is acceptable may still be an optimistic of the true level of beat tracking performance across all music.

5. DISCUSSION AND CONCLUSIONS

In order to estimate the confidence of a beat tracking output without ground truth data we propose the use of two methods based on the mutual agreement between a committee of beat tracking algorithms. The first, the Mean Mutual Agreement, was used for estimating the level of concensus between the beat outputs of the committee. The second, the Maximum Mutual Agreement, was used for selecting the best beat sequence from the committee of beat trackers.

Through a subjective listening test we determined an MMA threshold between this committee of beat trackers of 1.5 bits above which we believe automatic beat tracking can be applied with high confidence. Based on this perceptual confidence, we demonstrate that around 73% of the MillionSongSubset (and 73% of the MillionSongDataset) could be automatically annotated using our committee of beat trackers. This proportion of excerpts for which we can be confident in an automatic beat annotation was also verified in a second dataset with ground truth annotations. However, given the apparent bias in these datasets towards easier genres for beat tracking we consider the value of 73% to be somewhat optimistic. We plan to verify this hypothesis by measuring MMA in more diverse datasets.

Regarding the types of music which formed the remaining 27% on the MillionSongSubset (*i.e.*, those below the threshold) we found a high proportion of tags related to metal and similar "noisy" styles of music. Beyond classical music and jazz, which are known to be challenging for beat tracking systems, we consider the difficulty of beat tracking in metal to be a new (and unexpected) result, and furthermore an interesting area for the future development

of beat tracking algorithms.

In addition to using MMA to determine successful beat tracking, we also presented a related technique, MaxMA, to select beat estimations among a committee of beat trackers. The fact that a simple approach of this kind was able to demonstrate a significant improvement over using individual state of the art algorithms is encouraging. Yet, as our results indicate, performance of MaxMA falls some way below that of the Oracle system using our committee. This suggests that there is still room for making a more accurate selection among existing algorithms, and this will form a further avenue of future work.

One limitation of our approach may have been the use of short excerpts for the listening test. This was done to make the listening test as manageable as possible for a wide range of participants. However, to obtain a more comprehensive understanding of subjective ratings for longer musical excerpts and towards a better understanding of perceptual difficulty in beat perception we plan to conduct more sophisticated subjective listening experiments.

While all the directions for future work have so been related to beat tracking, we strongly believe that, given suitable evaluation metrics, our framework based on the MMA and MaxMA could be readily applied to other problems in MIR. We therefore encourage researchers to explore its usage in other MIR problems such as onset detection, chord detection, structural segmentation, and music transcription.

We also plan to continue addressing difficulty cases for beat tracking using the MMA in more diverse datasets and use the MaxMA approach to annotate large datasets. We will also attempt a more comprehensive listen test using longer audio excerpts, a more diverse variety of categories and genres to better quantify perceptually difficulty in beat perception.

6. ACKNOWLEDGEMENTS

Omitted.

7. REFERENCES

[1] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman and P. Lamere, "The Million Song Dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pp. 591–596, 2011.

[2] 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.

[3] 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.

[4] 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, pp. 290–301, 2012.

[5] S. Dixon, "Evaluation of the audio beat tracking system BeatRoot," *Journal of New Music Research*, Vol. 36, pp. 39–50, 2007.

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

[7] 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,* PhD. Thesis. Music Technology Group, Universitat Pompeu Fabra, 2005.

[8] 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.

[9] A. Holzapfel and Y. Stylianou "Parataxis: Morphological similarity in traditional music," *Proc. of the 11th ISMIR Conference,* pp. 453–458, 2010.

[10] Omitted for blind review

[11] 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

[12] 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.

[13] M. F. McKinney, D. Moelants, M. E. P. Davies, A. Klapuri, "Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms," *Journal of New Music Research,* Vol. 36, pp. 1–16, 2007.

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

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