ASSIGNING A CONFIDENCE THRESHOLD ON AUTOMATIC BEAT ANNOTATION IN LARGE DATASETS

José R. Zapata[1], André Holzapfel[1], Matthew E.P. Davies[2], João L.Oliveira[2,3] and Fabien Gouyon[2,3]
[1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2]Sound and Music Computing Group, INESC TEC, Porto, Portugal
[3]Faculty of Engineering of the University of Porto, Porto, Portugal
joser.zapata@upf.edu, hannover@csd.uoc.gr, {mdavies, jmso, fgouyon}@inescporto.pt

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 a 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 tracker outputs.

1. INTRODUCTION

Beat tracking can be considered one of the fundamental problems in music information retrieval (MIR) research. There have been numerous algorithms presented (*e.g.*, [5, 6, 10]) 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 [8], structural segmentation [11]) 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, 12]. Indeed the viewpoint could be taken that beat tracking performance is approaching a glass ceiling [9] with the current algorithms stagnating at around the 80% mark when evaluated using the least stringent metrics on common datasets [4].

In previous work [9] we proposed that the presence of this apparent glass ceiling was not the result of beat tracking algorithms having reached their full potential, but rather the datasets on which beat trackers are evaluated not containing a sufficient proportion of challenging examples; and that current beat trackers have over-learned the musical properties of the "easier" songs 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 [9]. Our technique was based on measuring the mean mutual agreement (MMA) between a committee of state of the art beat tracking algorithms, where low mutual agreement (or put another way, high disagreement) between beat outputs was shown to be a good indicator of low performance against the ground truth. To this end we empirically determined an MMA "failure" threshold below which beat tracking performance was shown to be very poor, and created a new database comprised of challenging songs with MMA below this threshold.

In this paper we address the opposite issue, where, 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 beat tracking performance 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 be overly optimistic. Therefore when seeking to determine an unbiased measure of performance we can either manually annotate more and more music examples for evaluation, or instead attempt to estimate beat tracking performance without ground truth. Due to the impractical nature of the first option, we pursue the second. Furthermore, if no ground truth is required, then performance can be estimated on very large (effectively unlimited) collections of music.

We extend our previous work to attempt to determine an MMA "success" threshold above which we can have high confidence in the beat tracking output of a commit-

tee of state of the art algorithms. 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 songs for which the MMA has been calculated. In each case the beat tracker output chosen to represent the committee is selected automatically as the one which most agrees with the remainder of the committee, *i.e.*, the beat tracker output with the maximal mutual agreement (MaxMA). We demonstrate that selecting between beat tracker outputs using MaxMA leads to improved performance over consistently 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 songs 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 the proposed method based on mutual agreement and describes the chosen committee. Section 3 demonstrates the improvement in performance when selecting a beat tracker based on the MaxMA approach on a manually annotated dataset. Section 4 applies the technique to non-annotated data and describes the procedure followed in the listening test and the main results. Section 5 concludes the paper with discussion of the results and areas for future work.

2. MEASURING MUTUAL AGREEMENT

The measurement of Mean Mutual Agreement (MMA) is inspired by the Query by Committee concept [14] which selects the most informative set of samples from a database based on the mutual (dis-)agreement between a designated committee of learners. In beat tracking, the MMA is computed using the beat outputs (or beat sequences) of a committee of N beat trackers on a musical piece, by measuring the mutual agreement $MA_{i,j}$ between every pair of estimated beat tracker outputs i and j, and retrieving the mean of all $N(N-1)/2$ mutual agreements. A graphical example is shown in Figure 1.

In addition to calculating the MMA as a summary statistic, we can easily identify the mutual agreement, MA_i, of the beat tracker output i which most agrees with the remainder of the committee: MaxMA, and the beat tracker output i which agrees the least: MinMA. In order to measure the mutual agreement $MA_{i,j}$ between each pair $\{i,j\}$ of beat tracker outputs, a beat tracking evaluation method must be chosen. In [9] we reviewed the properties of existing evaluation methods [2] and selected the Information Gain approach [3] (InfGain) as the only one with a true zero value, able to match low MMA (measured in bits) with unrelated beat tracker outputs:

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

The Information Gain measure is determined by forming a beat error histogram representing the timing error between beat sequences. 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, where the upper limit is $\log_2(K)$ for K=40 histogram bins. For further details see [3].

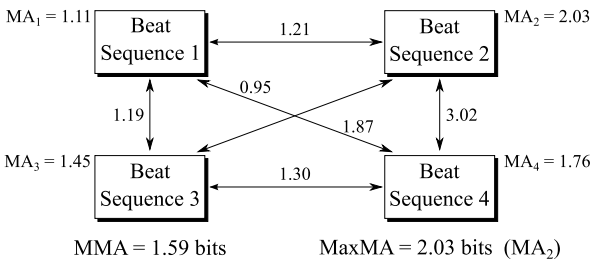

Figure 1: Example calculation of the MMA and MaxMA for a song with the beats estimated from a committee of four beat trackers.

To form our committee we select five state of the art and publicly available beat trackers: Dixon (Dix.) [5], Degara (Deg.) [4], Ellis (Ell.) [6], IBT [13], and Klapuri (Kla.) [10]. These convey the performance and diversity necessary to compute a reliable MMA [9].

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 tracker output we computed and compared the outputs of this committee on a manually annotated dataset containing 1360 song excerpts [5, 7] (referred to as **Dataset1360**) which covers the following genres: Acoustic; Afro-American; Jazz/Blues; Classical; Choral; Electronic; Rock/Pop; Balkan/Greek; and Samba.

Since we have shown in previous work that disagreement among the committee indicates poor beat tracking performance [9], we consider the potential positive effect of agreement within the committee. Our hypothesis is that the beat tracker that best agrees with the rest of the committee (the one with MaxMA) will be the most reliable algorithm for a specific musical piece. On this basis, we compare the mean ground truth performance of the best overall beat tracker, Best Mean, (which was shown to be Klapuri [10] (Kla.) for Dataset1360 [9]) against the mean scores of the algorithms with the MaxMA and MinMA for each excerpt. To illustrate the upper limit on performance for our committee we also compute the Oracle as the mean score given by the best beat tracker per excerpt.

Figure 2 compares the results of the described performance variants on Dataset1360. As described in Section 2, the MaxMA and MinMA were computed using the InfGain [1]. In order to compare MinMA and MaxMA against

[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`

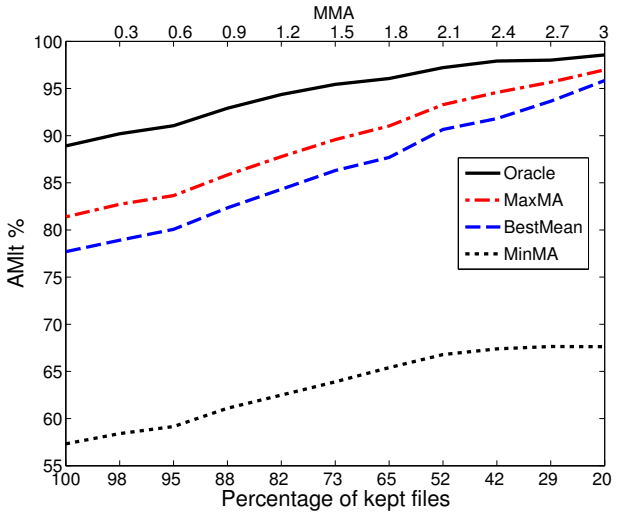

Figure 2: AMLt scores of the beat tracker output with maximum (MaxMA) and minimum (MinMA) agreement per song, compared with the single best beat tracker choice (BestMean), and the oracle score (Oracle) for various thresholds of MMA applied to Dataset1360.

the Best Mean and Oracle performances of the committee on the same data, we used the least stringent continuity-based measure, AMLt [1] (Allowed Metrical Level with no continuity required) [3], where beats are accurate when consecutive falling within tempo-dependent tolerance windows around successive annotations. Beat tracker outputs are also considered accurate if beats occur on the off-beat, or are estimated at double or half the annotated tempo. This performance measure provides a more intuitive scale of 0 to 100% than Information Gain and allows some ambiguity in the choice of metrical level at which the beats are estimated.

Performance across these conditions was 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 excerpts from the dataset (*e.g.*, for an MMA of 2.1 bits we retain 52.1% of the song in the dataset).

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 songs for beat tracking, and therefore can confine the data to easier songs by removing those with low MMA. Across all MMA thresholds we can observe that the performance of MinMA is significantly lower than all other configurations tested. Although lower than the Oracle, MaxMA outperforms the BestMean algorithm, and the difference between the two, around 3.3%, is statistically significantly ($p<0.01$) for all songs with an MMA below 2.4 bits. Above 2.4 bits this difference is no longer significant however the performance of the Oracle, BestMean and MaxMA are all very high. This suggests that for very high MMA thresholds, where beat tracker outputs are highly consistent with one another, any attempt to choose between the members of the committee offers little scope for improvement.

4. AUTOMATICALLY BEAT-ANNOTATING A LARGE DATASET

Having illustrated the validity of using the MaxMA method to select a beat tracker output among a committee of algorithms on a manually 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 automatically annotate beats in large collections we employ our MMA and MaxMA methods and attempt to determine the proportion of songs for which the beat estimates are acceptable via a subjective listening test. We want to establish a threshold on MMA above which the beat tracker outputs are perceptually acceptable. For each file, the beat tracker 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 songs without ground truth for which audio previews were obtained. The majority of audio previews were either 30 s or 60 s in duration, however to provide sufficiently long song excerpts for beat tracking we discarded any shorter than 20 s. This left a set of 9940 songs on which to automatically annotate beats. To complement the audio data, we obtained 31696 Last.fm [2] tags which covered 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 tracker output.

4.2 Subjective Listening Test

The aim of our listening test was to determine an MMA threshold above which the beat tracker output given by the MaxMA method was deemed acceptable to human listeners. By subsequent inspection of the number of songs in the dataset above this MMA threshold we could then 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 songs, it is equally impractical to ask participants to listen and rate this large number. As alternative to the exhaustive rating of all audio songs, we selected 8 levels of MMA = [0.5, 1.0, 1.5, ..., 4.0] bits and chose the 6 closest songs from the MillionSongSubset to each MMA level, giving a total of 48 songs 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 songs, which might bias the listener ratings, each musical stimulus was formed out of the middle 15 s of each

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

song. 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 an individual playlist of songs 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 were: 1 - Bad, 2 - Poor, 3 - Fair, 4 - Good, 5 - Excellent.

4.3 Results

4.3.1 Listening Test

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 rating was 3.7 (Good) with a standard deviation of 0.93. However, for MMA equal to 1 bit, the mean rating was much lower, at around 2.4 (Poor). Performing a t-test, we found the difference between the mean ratings at these MMA values to be highly significant ($p < 0.0001$). On this basis we can easily identify an MMA threshold of 1.5 bits which separates perceptually acceptable beat tracking from inaccurate beat tracking.

4.3.2 MMA Threshold

By selecting an MMA of 1.5 bits as a threshold of perceptual confidence for beat tracking we find 996 songs (73%) in Dataset1360 and 7252 songs (coincidentally also 73%), in the MillionSongSubset above this limit (see Figure 4). Table 1 shows the AMLt scores for the Oracle, MaxMA, Best Mean, and MinMA for the two subsets of Dataset1360 separated by MMA = 1.5 bits, evaluated against the ground truth. The beat tracking performance is consistently high for songs 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 the objective evaluation. While beat tracking performance is lower for MMA < 1.5 bits this does not mean the MaxMA beat estimations cannot be perceptually accurate, merely that we do not have high confidence in them.

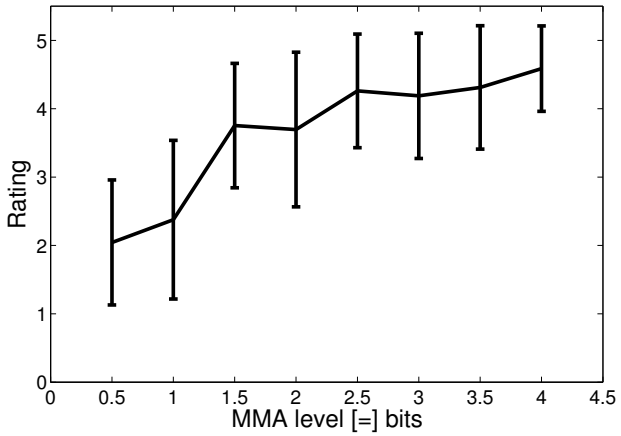

Figure 3: Listening test ratings *vs* MMA for the selected 48 music excerpts, from the MillionSongSubset.

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.

4.3.3 Last.fm Tag Analysis

Given the MMA threshold and collected Last.fm metadata, we now look at the genre-related tags of the songs 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 all genre labels related to metal music below the threshold was a surprising result since this music is strongly percussive and is not 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 we are unaware of many metal examples in existing beat tracking databases. This suggests 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

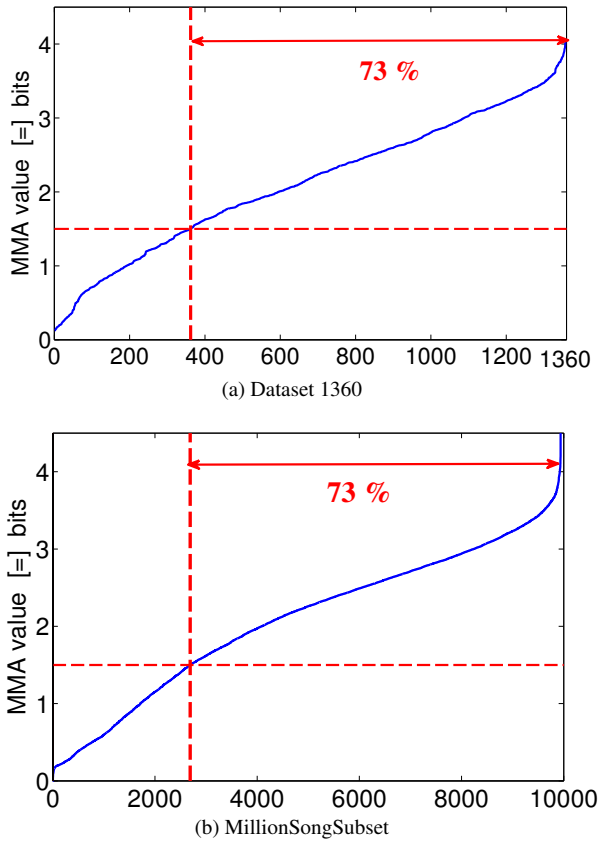

Figure 4: Datasets sorted by MMA and the perceptual threshold of 1.5 bits.

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.

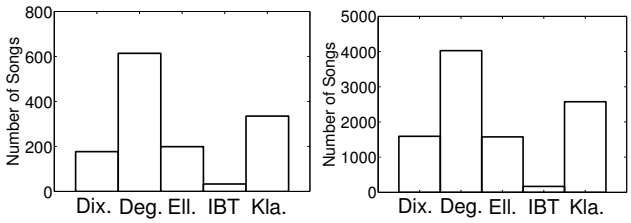

Figure 5: Histograms with the number of times each algorithm is chosen with the MaxMA approach.

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 [9], 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 songs for beat tracking in this dataset, our estimate of 73% of songs for which beat tracking is acceptable may still be an optimistic estimate of the true level of beat tracking performance across all music.

4.3.4 MaxMA Choice of Beat Tracker

Having investigated the main results of applying MaxMA to automatically annotate beat locations, we now address the properties of the committee. Figure 5 presents histograms for both evaluated datasets depicting the proportion of songs where each beat tracking algorithm is selected as the MaxMA beat output. Both histograms show similar shapes, indicating that there may be some similar properties between the musical content of both datasets. The two most chosen algorithms are those of Degara [4] and Klapuri [10]; both of which perform most accurately against the ground truth, and can be considered the best among the state of the art methods. As to why the Degara algorithm is chosen more frequently than that of Klapuri, results in [4] indicate that the inter-quartile range of the Degara algorithm is smaller than that of Klapuri (for a similar median), implying it is "wrong" in a lower proportion of songs.

5. DISCUSSION AND CONCLUSIONS

To estimate the confidence of beat tracking without ground truth annotations we have proposed 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 to estimate the level of consensus between the beat outputs of the committee. The second, the Maximum Mutual Agreement, was used for selecting the best beat tracking output 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 could be automatically annotated using our committee of beat trackers. This proportion of songs for which we can be confident in an automatic beat annotation was also verified in a second dataset with manually annotated ground truth. Given the apparent bias in

these datasets towards easier genres for beat tracking, we consider this value of 73% to be somewhat optimistic. We plan to verify this hypothesis in future work by measuring MMA in more diverse datasets.

Regarding the types of music which formed the remaining 27% of 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 exploring new selection methods will form a further area for future work.

One limitation of our approach may have been the use of short song 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 greater understanding of subjective ratings for longer musical excerpts and 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 far been related to beat tracking, we strongly believe that, given suitable evaluation metrics, our framework based on MMA and MaxMA could be readily applied to other areas of MIR. We therefore encourage researchers to explore its usage in problems such as onset detection, chord detection, structural segmentation, and music transcription.

6. ACKNOWLEDGEMENTS

This research received support from the Portuguese Foundation for Science and Technology through the "ShakeIt" project (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.

7. REFERENCES

[1] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman and P. Lamere, "The Million Song Dataset," in *Proc. of the 12th ISMIR conference*, 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. MTG, Universitat Pompeu Fabra, 2005.

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

[9] A. Holzapfel, M. E. P. Davies, J.R. Zapata, J.L. Oliveira and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Transactions on Audio, Speech and Language Processing,* In press, 2012.

[10] 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

[11] 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.

[12] 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.

[13] 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.

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