EVALUATING THE EVALUATION MEASURES FOR BEAT TRACKING

Mathew E. P. Davies
Sound and Music Computing Group
INESC TEC, Porto, Portugal
mdavies@inesctec.pt

Sebastian Böck
Department of Computational Perception
Johannes Kepler University, Linz, Austria
sebastian.boeck@jku.at

ABSTRACT

The evaluation of audio beat tracking systems is normally addressed in one of two ways. One approach is for human listeners to judge performance by listening to beat times mixed as clicks with music signals. The more common alternative is to compare beat times against ground truth annotations via one or more of the many objective evaluation measures. However, despite a large body of work in audio beat tracking, there is currently no consensus over which evaluation measure(s) to use, meaning multiple accuracy scores are typically reported. In this paper, we seek to evaluate the evaluation measures by examining the relationship between objective accuracy scores and human judgements of beat tracking performance. First, we present the raw correlation between objective scores and subjective ratings, and show that evaluation measures which allow alternative metrical levels appear more correlated than those which do not. Second, we explore the effect of parameterisation of objective evaluation measures, and demonstrate that correlation is maximised for smaller tolerance windows than those currently used. Our analysis suggests that true beat tracking performance is currently being overestimated via objective evaluation.

1. INTRODUCTION

Evaluation is a critical element of music information retrieval (MIR) [16]. Its primary use is a mechanism to determine the individual and comparative performance of algorithms for given MIR tasks towards improving them in light of identified strengths and weaknesses. Each year many different MIR systems are formally evaluated within the MIREX initiative [6].

In the context of beat tracking, the concept and purpose of evaluation can be addressed in several ways. For example, to measure reaction time across changing tempi [2], to identify challenging musical properties for beat trackers [9] or to drive the composition of new test datasets [10]. However, as with other MIR tasks, evaluation in beat tracking is most commonly used to estimate the performance of one or more algorithms on a test dataset.

© Mathew E. P. Davies, Sebastian Böck.
Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Mathew E. P. Davies, Sebastian Böck. "Evaluating the evaluation measures for beat tracking", 15th International Society for Music Information Retrieval Conference, 2014.

This measurement of performance can happen via subjective listening test, where human judgements are used to determine beat tracking performance [3], to discover: *how perceptually accurate the beat estimates are when mixed with the input audio*. Alternatively, objective evaluation measures can be used to compare beat times with ground truth annotations [4], to determine: *how consistent the beat estimates are with the ground truth according to some mathematical relationship*. While undertaking listening tests and annotating beat locations are both extremely time-consuming tasks, the apparent advantage of the objective approach is that once ground truth annotations have been determined, they can easily be re-used without the need for repeated listening experiments. However, the usefulness of any given objective accuracy score (of which there are many [4]) is contingent on its ability to reflect human judgement of beat tracking performance. Furthermore, for the entire objective evaluation process to be meaningful, we must rely on the inherent accuracy of the ground truth annotations.

In this paper we work under the assumption that musically trained experts can provide meaningful ground truth annotations and rather focus on the properties of the objective evaluation measures. The main question we seek to address is: *to what extent do existing objective accuracy scores reflect subjective human judgement of beat tracking performance?* In order to answer this question, even in principle, we must first verify that human listeners can make reliable judgements of beat tracking performance. While very few studies exist, we can find supporting evidence suggesting human judgements of beat tracking accuracy are highly repeatable [3] and that human listeners can reliably disambiguate accurate from inaccurate beat click sequences mixed with music signals [11].

The analysis we present involves the use of a test database for which we have a set of estimated beat locations, annotated ground truth and human subjective judgements of beat tracking performance. Access to all of these components (via the results of existing research [12, 17]) allows us to examine the correlation between objective accuracy scores, obtained by comparing the beat estimates to the ground truth, with human listener judgements. To the best of our knowledge this is the first study of this type for musical beat tracking.

The remainder of this paper is structured as follows. In Section 2 we summarise the objective beat tracking evaluation measures used in this paper. In Section 3 we describe

the comparison between subjective ratings and objective scores of beat tracking accuracy. Finally, in Section 4 we present discussion and areas for future work.

2. BEAT TRACKING EVALUATION MEASURES

In this section we present a brief summary each of the evaluation measures from [4]. While nine different approaches were presented in [4], we reduce them to seven by only presenting the underlying approaches for comparing a set of beats with a set of annotations (i.e. ignoring alternate metrical interpretations). We consider the inclusion of different metrical interpretations of the annotations to be a separate process which can be applied to any of these evaluation measures (as in [5, 8, 15]), rather than a specific property of one particular approach. To this end, we choose three evaluation conditions: *Annotated* – comparing beats to annotations, *Annotated+Offbeat* – including the "off-beat" of the annotations for comparison against beats and *Annotated+Offbeat+D/H* – including the off-beat and both double and half the tempo of the annotations. This doubling and halving has been commonly used in beat tracking evaluation to attempt to reflect the inherent ambiguity in music over which metrical level to tap the beat [13]. The set of seven basic evaluation measures are summarised below:

F-measure : accuracy is determined through the proportion of *hits*, *false positives* and *false negatives* for a given annotated musical excerpt, where *hits* count as beat estimates which fall within a pre-defined tolerance window around individual ground truth annotations, *false positives* are extra beat estimates, and *false negatives* are missed annotations. The default value for the tolerance window is $\pm$0.07s.

PScore : accuracy is measured as the normalised sum of the cross-correlation between two impulse trains, one corresponding to estimated beat locations, and the other to ground truth annotations. The cross-correlation is limited to the range covering 20% of the median inter-annotation-interval (IAI).

Cemgil : a Gaussian error function is placed around each ground truth annotation and accuracy is measured as the sum of the "errors" of the closest beat to each annotation, normalised by whichever is greater, the number of beats or annotations. The standard deviation of this Gaussian is set at 0.04s.

Goto : the annotation interval-normalised timing error is measured between annotations and beat estimates, and a binary measure of accuracy is determined based on whether a region covering 25% of the annotations continuously meets three conditions – the maximum error is less than $\pm$17.5% of the IAI, and the mean and standard deviation of the error are within $\pm$10% of the IAI.

Continuity-based : a given beat is considered accurate if it falls within a tolerance window placed around an annotation and that the previous beat also falls within the preceding tolerance window. In addition, a separate threshold requires that the estimated inter-beat-interval should be close to the IAI. In practice both thresholds are set at $\pm$17.5% of the IAI. In [4], two basic conditions consider the ratio of the longest continuously correct region to the length of the excerpt (CMLc), and the total proportion of correct regions (CMLt). In addition, the AMLc and AMLt versions allow for additional interpretations of the annotations to be considered accurate. As specified above, we reduce these four to two principal accuracy scores. To prevent any ambiguity, we rename these accuracy scores **Continuity-C** (CMLc) and **Continuity-T** (CMLt).

Information Gain : this method performs a two-way comparison of estimated beat times to annotations and vice-versa. In each case, a histogram of timing errors is created and from this the **Information Gain** is calculated as the Kullback-Leibler divergence from a uniform histogram. The default number of bins used in the histogram is 40.

3. SUBJECTIVE VS. OBJECTIVE COMPARISON

3.1 Test Dataset

To facilitate the comparison of objective evaluation scores and subjective ratings we require a test dataset of audio examples for which we have both annotated ground truth beat locations and a set of human judgements of beat tracking performance for a beat tracking algorithm. For this purpose we use the test dataset from [17] which contains 48 audio excerpts (each 15s in duration). The excerpts were selected from the MillionSongSubset [1] according to a measurement of mutual agreement between a committee of five state of the art beat tracking algorithms. They cover a range from very low mutual agreement – shown to be indicative of beat tracking difficulty, up to very high mutual agreement – shown to be easier for beat tracking algorithms [10].

In [17] a listening experiment was conducted where a set of 22 participants listened to these audio examples mixed with clicks corresponding to automatic beat estimates and rated on a 1 to 5 scale how well they considered the clicks represented the beats present in the music. For each excerpt these beat times were the output of the beat tracker which most agreed with the remainder of the five committee members from [10]. Analysis of the subjective ratings and measurements of mutual agreement revealed low agreement to be indicative of poor subjective performance.

In a later study, these audio excerpts were used as one test set in a beat tapping experiment, where participants tapped the beat using a custom piece of software [12]. In order to compare the mutual agreement between tappers with their global performance against the ground truth, a musical expert annotated ground truth beat locations. The tempi range from 62 BPM (beats per minute) up to 181 BPM and, with the exception of two excerpts, all are in 4/4 time. Of the remaining two excerpts, one is in 3/4 time and

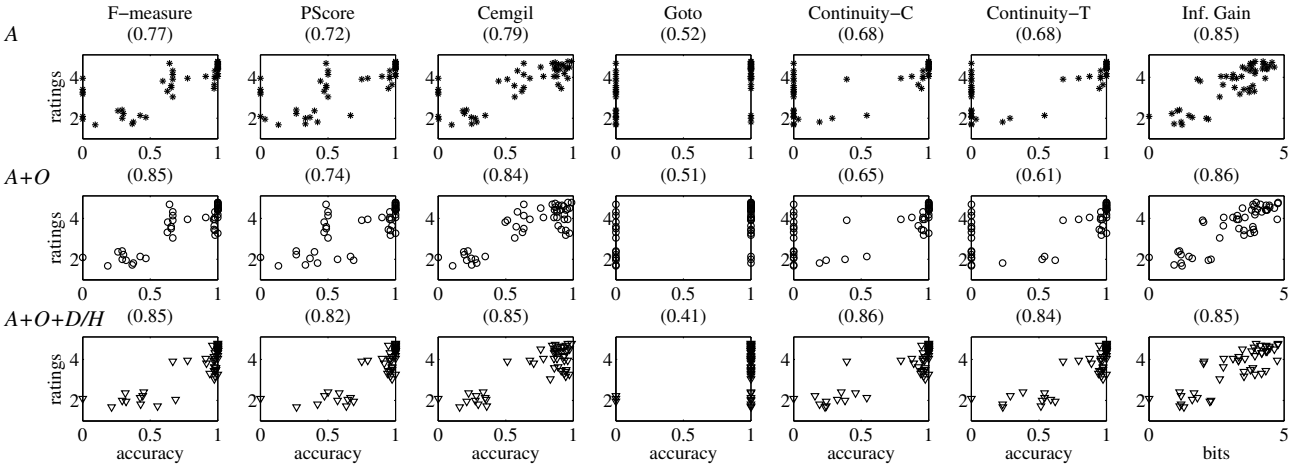

Figure 1. Subjective ratings vs. objective accuracy scores for different evaluation measures. The rows indicate different evaluation conditions. (top row) *Annotated*, (middle row) *Annotated+Offbeat*, and (bottom row) *Annotated+Offbeat+D/H*. For each scatter plot, the linear correlation coefficient is provided.

the other was deemed to have no beat at all, and therefore no beats were annotated.

In the context of this paper, this set of ground truth beat annotations provides the final element required to evaluate the evaluation measures, since we now have: i) automatically estimated beat locations, ii) subjective ratings corresponding to these beats and iii) ground truth annotations to which the estimated beat locations can be compared. We use each of the seven evaluation measures described in Section 2 to obtain the objective accuracy scores according to the three versions of the annotations: *Annotated*, *Annotated+Offbeat* and *Annotated+Offbeat+D/H*. Since all excerpts are short, and we are evaluating the output of an offline beat tracking algorithm, we remove the startup condition from [4] where beat times in the first five seconds are ignored.

3.2 Results

3.2.1 Correlation Analysis

To investigate the relationship between the objective accuracy scores and subjective ratings, we present scatter plots in Figure 1. The title of each individual scatter plot includes the linear correlation coefficient which we interpret as an indicator of the validity of a given evaluation measure in the context of this dataset.

The highest overall correlation (0.86) occurs for **Continuity-C** when the offbeat and double/half conditions are included. However, for all but **Goto**, the correlation is greater than 0.80 once these additional evaluation criteria are included. It is important to note only **Continuity-C** and **Continuity-T** explicitly include these conditions in [4]. Since **Goto** provides a binary assessment of beat tracking performance, it is unlikely to be highly correlated with the subjective ratings from [17] where participants were explicitly required to use a five point scale rather than a good/bad response concerning beat tracking performance. Nevertheless, we retain it to maintain consistency with [4].

Comparing each individual measure across these evaluation conditions, reveals that **Information Gain** is least affected by the inclusion of additional interpretations of the annotations, and hence most robust to ambiguity over metrical level. Referring to the **F-measure** and **PScore** columns of Figure 1 we see that the "vertical" structure close to accuracies of 0.66 and 0.5 respectively is mapped across to 1 for the *Annotated+Offbeat+D/H* condition. This pattern is also reflected for **Goto**, **Continuity-C** and **Continuity-T** which also determine beat tracking accuracy according to fixed tolerance windows, i.e. a beat falling anywhere inside a tolerance window is perfectly accurate. However, the fact that a fairly uniform range of subjective ratings between 3 and 5 (i.e. "fair" to "excellent" [17]) exists for apparently perfect objective scores indicates a potential mismatch and over-estimation of beat tracking accuracy. While a better visual correlation appears to exist in the scatter plots of **Cemgil** and **Information Gain**, this is not reflected in the correlation values (at least not for the *Annotated+Offbeat+D/H* condition). The use a Gaussian instead of a "top-hat" style tolerance window for **Cemgil** provides more information regarding the precise localisation of beats to annotations and hence does not have this clustering at the maximum performance. The **Information Gain** measure does not use tolerance windows at all, instead it measures beat tracking accuracy in terms of the temporal dependence between beats and annotations, and thus shows a similar behaviour.

3.2.2 The Effect of Parameterisation

For the initial correlation analysis, we only considered the default parameterisation of each evaluation measure as specified in [4]. However, to only interpret the validity of the evaluation measures in this way presupposes that they have already been optimally parameterised. We now explore whether this is indeed the case, by calculating the objective accuracy scores (under each evaluation condition) as a function of a threshold parameter for each measure.

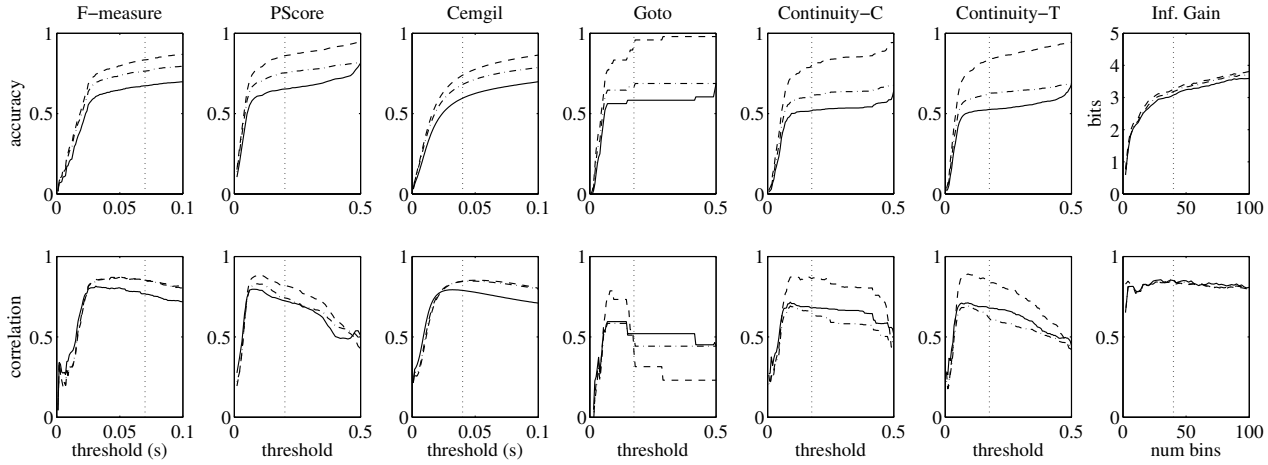

Figure 2. (top row) Beat tracking accuracy as a function of threshold (or number of bins for Information Gain) per evaluation measure. (bottom row) Correlation between subjective ratings and accuracy scores as a function of threshold (or number of bins). In each plot the solid line indicates the *Annotated* condition, the dashed–dotted line shows *Annotated+Offbeat* and the dashed line shows *Annotated+Offbeat+D/H*. For each evaluation measure, the default parameteristation from [4] is shown by a dotted vertical line.

We then re-compute the subjective vs. objective correlation. We adopt the following parameter ranges as follows:

F-measure : the size of the tolerance window increases from $\pm$0.001s to $\pm$0.1s.

PScore : the width of the cross-correlation increases from 0.01 to 0.5 times the median IAI.

Cemgil : the standard deviation of the Gaussian error function grows from 0.001s to 0.1s.

Goto : to allow a similar one-dimensional representation, we make all three parameters identical and vary them from $\pm$0.005 to $\pm$0.5 times the IAI.

Continuity-based : the size of the tolerance window increases from $\pm$0.005 to $\pm$0.5 times the IAI.

Information Gain : we vary the number of bins in multiples of 2 from 2 up to 100.

In the top row of Figure 2 the objective accuracy scores as a function of different parameterisations are shown. The plots in the bottom row show the corresponding correlations with subjective ratings. In each plot the dotted vertical line indicates the default parameters. From the top row plots we can observe the expected trend that, as the size of the tolerance window increases so the objective accuracy scores increase. For the case of **Information Gain** the beat error histograms become increasingly sparse due to having more histogram bins than observations, hence the entropy reduces and the information gain increases. In addition, **Information Gain** does not have a maximum value of 1, but instead, $\log_2$ of the number of histogram bins [4].

Looking at the effect of correlation with subjective ratings in the bottom row of Figure 2, we see that for most evaluation measures there is rapid increase in the correlation as the tolerance windows grow from very small sizes after which the correlation soon reaches its maximum and then reduces. Comparing these change points with the dotted vertical lines (which show the default parameters) we see that correlation is maximised for smaller (i.e. more restrictive) parameters than those currently used. By finding the point of maximum correlation in each of the plots in the bottom row of Figure 2 we can identify the parameters which yield the highest correlation between objective accuracy and subjective ratings. These are shown for the *Annotated+Offbeat+D/H* evaluation condition in Table 1 for which the correlation is typically highest. Returning to the plots in the top row of Figure 2 we can then read off the corresponding objective accuracy with the default and then maximum correlation parameters. These accuracy scores are shown in Table 2.

	Default Parameters	Max. Correlation Parameters
F-measure	0.070s	0.049s
PScore	0.200	0.110
Cemgil	0.040s	0.051s
Goto	0.175	0.100
Continuity-C	0.175	0.095
Continuity-T	0.175	0.090
Information Gain	40	38

Table 1. Comparison of default parameters per evaluation measure with those which provide the maximum correlation with subjective ratings in the *Annotated+Offbeat+D/H* condition.

From these Tables we see that it is only **Cemgil** whose default parameterisation is lower than that which maximises the correlation. However this does not apply for the *Annotated* only condition which is implemented in [4]. While there is a small difference for **Information Gain**, in-

	Annotated		*Annotated+Offbeat*		*Annotated+Offbeat+D/H*	
	Default Params	Max Corr. Params	Default Params	Max Corr. Params	Default Params	Max Corr. Params
F-measure	0.673	0.607	0.764	0.738	0.834	0.797
PScore	0.653	0.580	0.753	0.694	0.860	0.792
Cemgil	0.596	0.559	0.681	0.702	0.739	0.779
Goto	0.583	0.563	0.667	0.646	0.938	0.813
Continuity-C	0.518	0.488	0.605	0.570	0.802	0.732
Continuity-T	0.526	0.505	0.624	0.587	0.837	0.754
Information Gain	3.078	2.961	3.187	3.187	3.259	3.216

Table 2. Summary of objective beat tracking accuracy under the three evaluation conditions: *Annotated*, *Annotated+Offbeat* and *Annotated+Offbeat+D/H* per evaluation measure. Accuracy is reported using the default parameterisation from [4] and also using the parameterisation which provides maximal correlation to the subjective ratings. For **Information Gain** only performance is measured in bits.

spection of Figure 2 shows that it is unaffected by varying the number of histogram bins in terms of the correlation. In addition, the inclusion of the extra evaluation criteria also leads to a negligible difference in reported accuracy. Therefore **Information Gain** is most robust to parameter sensitivity and metrical ambiguity. For the other evaluation measures the inclusion of the *Annotated+Offbeat* and the *Annotated+Offbeat+D/H* (in particular) leads to more pronounced differences. The highest overall correlation between objective accuracy scores and subjective ratings (0.89) occurs for **Continuity-T** for a tolerance window of $\pm 9\%$ of the IAI rather than the default value of $\pm 17.5\%$. Referring again to Table 2 we see that this smaller tolerance window causes a drop in reported accuracy from 0.837 to 0.754. Indeed a similar drop in performance can be observed for most evaluation measures.

4. DISCUSSION

Based on the analysis of objective accuracy scores and subjective ratings on this dataset of 48 excerpts, we can infer that: i) a higher correlation typically exists when the *Annotated+Offbeat* and/or *Annotated+Offbeat+D/H* conditions are included, and ii) for the majority of existing evaluation measures, this correlation is maximised for a more restrictive parameterisation than the default parameters which are currently used [4]. A strict following of the results presented here would promote either the use of **Continuty-T** for the *Annotated+Offbeat+D/H* condition with a smaller tolerance window, or **Information Gain** since it is most resilient to these variable evaluation conditions while maintaining a high subjective vs. objective correlation.

If we are to extrapolate these results to all existing work in the beat tracking literature this would imply that any papers reporting only performance for the *Annotated* condition using **F-measure** and **PScore** may not be as representative of subjective ratings (and hence true performance) as they could be by incorporating additional evaluation conditions. In addition, we could infer that most presented accuracy scores (irrespective of evaluation measure or evaluation condition) are somewhat inflated due to the use of artificially generous parameterisations. On this basis, we might argue that the apparent glass ceiling of around 80% for beat tracking [10] (using **Continuity-T** for the *Annotated+Offbeat+D/H* condition) may in fact be closer to 75%, or perhaps lower still. In terms of external evidence to support our findings, a perceptual study evaluating human tapping ability [7] used a tolerance window of $\pm 10\%$ of the IAI, which is much closer to our "maximum correlation" **Continuity-T** parameter of $\pm 9\%$ than the default value of $\pm 17.5\%$ of the IAI.

Before making recommendations to the MIR community with regard to how beat tracking evaluation should be conducted in the future, we should first revisit the makeup of the dataset to assess the scope from which we can draw conclusions. All excerpts are just 15s in duration, and therefore not only much shorter than complete songs, but also significantly shorter than most annotated excerpts in existing datasets (e.g. 40s in [10]). Therefore, based on our results, we cannot yet claim that our subjective vs. objective correlations will hold for evaluating longer excerpts. We can reasonably speculate that an evaluation across overlapping 15s windows could provide some local information about beat tracking performance for longer pieces, however this is currently not how beat tracking evaluation is addressed. Instead, a single score of accuracy is normally reported regardless of excerpt length. With the exception of [3] we are unaware of any other research where subjective beat tracking performance has been measured across full songs.

Regarding the composition of our dataset, we should also be aware that the excerpts were chosen in an unsupervised data-driven manner. Since they were sampled from a much larger collection of excerpts [1] we do not believe there is any intrinsic bias in their distribution other than any which might exist across the composition of the MillionSongSubset itself. The downside of this unsupervised sampling is that we do not have full control over exploring specific interesting beat tracking conditions such as off-beat tapping, expressive timing, the effect of related metrical levels and non-4/4 time-signatures. We can say that for the few test examples where the evaluated beat tracker tapped the off-beat (shown as zero accuracy points in the *Anno-*

tated condition but non-zero for the *Annotated+Offbeat* condition in Figure 1), were not rated as "bad". Likewise, there did not appear to be a strong preference over a single metrical level. Interestingly, the ratings for the *unannotatable* excerpt were among the lowest across the dataset.

Overall, we consider this to be a useful pilot study which we intend to follow up in future work with a more targeted experiment across a much larger musical collection. In addition, we will also explore the potential for using bootstrapping measures from Text-IR [14] which have also been used for the evaluation of evaluation measures. Based on these outcomes, we hope to be in a position to make stronger recommendations concerning how best to conduct beat tracking evaluation, ideally towards a single unambiguous measurement of beat tracking accuracy. However, we should remain open to the possibility that different evaluation measures may be more appropriate than others and that this could depend on several factors, including: the goal of the evaluation; the types of beat tracking systems evaluated; how the ground truth was annotated; and the make up of the test dataset.

To summarise, we believe the main contribution of this paper is to further raise the profile and importance of evaluation in MIR, and to encourage researchers to more strongly consider the properties of evaluation measures, rather than merely reporting accuracy scores and assuming them to be valid and correct. If we are to improve underlying analysis methods through iterative evaluation and refinement of algorithms, it is critical to optimise performance according to meaningful evaluation methodologies targeted towards specific scientific questions.

While the analysis presented here has only been applied in the context of beat tracking, we believe there is scope for similar subjective vs. objective comparisons in other MIR topics such as chord recognition or structural segmentation, where subjective assessments should be obtainable via similar listening experiments to those used here.

5. ACKNOWLEDGMENTS

This research was partially funded by the Media Arts and Technologies project (MAT), NORTE-07-0124-FEDER-000061, financed by the North Portugal Regional Operational Programme (ON.2–O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT) as well as FCT post-doctoral grant SFRH/BPD/88722/2012. It was also supported by the European Union Seventh Framework Programme FP7 / 2007-2013 through the GiantSteps project (grant agreement no. 610591).

6. REFERENCES

[1] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of 12th International Society for Music Information Retrieval Conference*, pages 591–596, 2011.

[2] N. Collins. *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*. PhD thesis, Centre for Music and Science, Faculty of Music, Cambridge University, 2006.

[3] R. B. Dannenberg. Toward automated holistic beat tracking, music analysis, and understanding. In *Proceedings of 6th International Conference on Music Information Retrieval*, pages 366–373, 2005.

[4] M. E. P. Davies, N. Degara, and M. D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. Technical Report C4DM-TR-09-06, Queen Mary University of London, Centre for Digital Music, 2009.

[5] S. Dixon. Evaluation of audio beat tracking system beatroot. *Journal of New Music Research*, 36(1):39–51, 2007.

[6] J. S. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[7] C. Drake, A. Penel, and E. Bigand. Tapping in time with mechanically and expressively performed music. *Music Perception*, 18(1):1–23, 2000.

[8] M. Goto and Y. Muraoka. Issues in evaluating beat tracking systems. In *Working Notes of the IJCAI-97 Workshop on Issues in AI and Music - Evaluation and Assessment*, pages 9–16, 1997.

[9] P. Grosche, M. Müller, 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*, pages 649–654, 2010.

[10] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. Oliveira, and F. Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(9):2539–2460, 2012.

[11] J. R. Iversen and A. D. Patel. The beat alignment test (BAT): Surveying beat processing abilities in the general population. In *Proceedings of the 10th International Conference on Music Perception and Cognition*, pages 465–468, 2008.

[12] M. Miron, F. Gouyon, M. E. P. Davies, and A. Holzapfel. Beat-Station: A real-time rhythm annotation software. In *Proceedings of the Sound and Music Computing Conference*, pages 729–734, 2013.

[13] D. Moelants and M. McKinney. Tempo perception and musical content: what makes a piece fast, slow or temporally ambiguous? In *Proceedings of the 8th International Conference on Music Perception and Cognition*, pages 558–562, 2004.

[14] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the International ACM SIGIR conference on research and development in information retrieval*, pages 525–532, 2006.

[15] A. M. Stark. *Musicians and Machines: Bridging the Semantic Gap in Live Performance*. PhD thesis, Centre for Digital Music, Queen Mary University of London, 2011.

[16] J. Urbano, M. Schedl, and X. Serra. Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, 41(3):345–369, 2013.

[17] J. R. Zapata, A. Holzapfel, M. E. P. Davies, J. L. Oliveira, and F. Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In *Proceedings of 13th International Society for Music Information Retrieval Conference*, pages 157–162, 2012.