ON THE USE OF ENTROPY FOR BEAT TRACKING EVALUATION

M. E. P. Davies and M. D. Plumbley

Queen Mary, University of London
Centre for Digital Music
Mile End Road, London E1 4NS, UK

ABSTRACT

Despite continued attention toward the problem of automatic beat detection in musical audio, the issue of how to evaluate beat tracking systems remains pertinent and controversial. As yet no consistent evaluation metric has been adopted by the research community. To this aim, we propose a new method for beat tracking evaluation by measuring beat accuracy in terms of the entropy of a beat error histogram. We demonstrate the ability of our approach to address several shortcomings of existing methods.

Index Terms— music, information retrieval, entropy

1. INTRODUCTION

The task of beat tracking is well known among researchers in music information retrieval. The common and simple analogy is that of foot-tapping in time to music. It is often reported that this seemingly trivial task for humans remains a significant challenge for computer systems [1]. The need for evaluation is simple; without it, we have no means to distinguish between good and bad cases, nor any way to gauge the performance of different beat tracking systems.

Human foot-tapping, although regarded as intuitive, is not always consistent [2]. Inconsistency in tapping can be described in several ways: most commonly, this refers to the metrical level at which the beats are tapped, e.g. for a given piece, some people will tap twice as fast as others; tapping may also occur at the same tempo, but in anti-phase i.e. on the *off-beat*; and thirdly, the localisation of the beats may not be precise, where some beats may be deemed to be ahead or behind the beat. Given the inherent ambiguity in the task, it is not surprising to discover that beat tracking algorithms are equally inconsistent in their behaviour. Despite the apparent simplicity of beat tracking, the task of measuring beat accuracy is complex and as yet no agreed upon method currently exists.

A straightforward form of beat tracking evaluation is the subjective listening test. Given a piece of music and a sequence of estimated beat locations, the beats can be rendered as percussive clicks which are then auditioned by human subjects who determine the accuracy of the tracking. Informally, this may be little more than a response to the query: 'do the beats sound in time?'. Dannenberg [3] performs subjective evaluation and defines correct tapping to be at twice or half the perceived metrical level, or on the off-beat. He also cites a specific case of incorrect tapping, known as *tempo drift*, where the beats are tapped at a slightly incorrect tempo and drift in and out of phase, which he labels as a perceptually disturbing error.

It is hard to dispute the validity of using human judgement to determine the accuracy of a perceptual construct such as beat location. However, subjective testing is time consuming, and this inherently limits the size of the test database upon which a beat tracker can be analysed. The alternative to subjective evaluation is pursue an objective approach. For each test example, this requires the annotation of ground truth beat locations against which the beat tracker output can be compared. For real musical performances (which are not performed in perfectly quantised time) hand annotation of beat locations is an extremely arduous task [4], which requires continued re-adjustment of beat locations often using audio and visual cues to obtain perceptually accurate data.

Given a ground truth sequence of beat annotations, the majority of existing evaluation methods define an allowance window around each annotation, and deem an individual beat to be correct if it occurs within this window. The thresholds defining the allowance window can either be fixed in time, e.g. 70ms [1] or tempo-dependent, e.g. 17.5% of the current inter-annotation-interval [5, 6]. Allowances can be made for the continuity of correct beats [5], or the beats may be treated as isolated events [1, 7]. However, the size of the allowance window is typically arbitrary. Depending on the thresholds used, the relative performance between competing algorithms can change [8]. Therefore a single window may be insufficient to obtain a complete picture of beat tracking performance.

To contend with ambiguity in tapping, either multiple phases and metrical levels of annotation are required, or the annotations must be re-sampled to each particular case. Inherently, this requires knowing the appropriate allowed metrical levels. Merely assuming that twice or half the rate will be acceptable may not generalise to all music, particularly those which do not have a 4/4 time-signature. Multiple passes over the annotation data can lead to multiple measures of accuracy, which require more interpretation than a single, all-encompassing accuracy value. The challenge is to meaningfully combine the different dimensions of metrical level, phase and localisation onto one single dimension of global beat accuracy.

In contrast to existing evaluation methods which model beat accuracy using allowance windows, we extract the timing error between beats and annotations. By analysing the error between all beats and annotations over annotation-centred beat-long windows we can analyse beat error independent of metrical level. We form a beat error histogram, where peaks represent implicit dependency between beats and annotations. We then define a measure of beat accuracy in terms of the entropy of the error distribution. We compare the properties between several evaluation methods, and demonstrate how our approach can address some of the shortcomings of existing methods.

The remainder of this paper is structured as follows. In section 2 we describe the approach for measuring beat tracking accuracy, followed in section 3 with some preliminary results and conclusions in section 4.

This research has been partially funded by EPSRC grants GR/S75802/01 and GR/S82213/01.

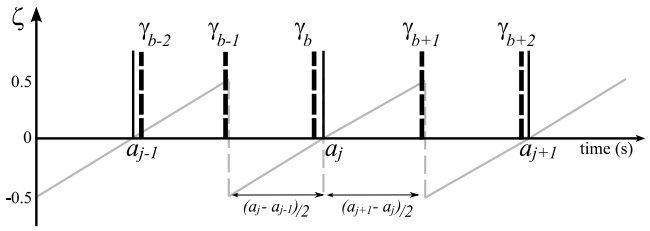

Fig. 1. Extraction of beat error ζ from beats γ_b and annotations a_j

2. APPROACH

In our objective approach to beat tracking evaluation we compare a sequence of extracted beat times γ to a sequence of ground truth beat annotations a. The basis of our method is the generation of a beat error histogram, which we can use to infer the relationship between the beats and the annotations. We can consider two such example histograms: i) a Dirac-delta distribution with an impulse at zero error, from which we should infer the beats were identical to the annotations; and ii) a uniform distribution, from which we should infer that the beats are entirely unrelated to the annotations. These examples merely represent the theoretical best and worst cases. For real beat tracking data, we can expect some form of distribution between these two extrema.

2.1. Beat error

To obtain the beat error data required to populate a histogram, we partition the input signal into beat length windows, centred on each annotation a_j. Within each annotation-centred window we could extract the time between the closest beat γ_b and a_j, however this would limit our analysis to the annotated metrical level [8]. Proceeding in this manner would also leave our approach blind to the trivial case of over-detection, where beats are placed at all time instants, so that for every annotation, some beat will exist with zero error.
To prevent the problem of over-detection and maintain analysis over multiple metrical levels we find the time between all beats γ_q that occur within each annotation window.

$$\gamma_q = \gamma_b \quad : \quad a_j - \Delta^*_{j-1} \leq \gamma_b < a_j + \Delta^*_j \tag{1}$$

To remove any dependency on the tempo of the input, we normalise the beat error $\zeta(j,q)$, for the q^{th} beat in the j^{th} annotation window, into the range [-0.5,0.5] beats by dividing the error by half the width of the annotation window, as shown in fig. 1

$$\zeta(j,q) = \begin{cases} \frac{\gamma_q - a_j}{\Delta^*_{j-1}} & \gamma_q \leq a_j \\ \\ \frac{\gamma_q - a_j}{\Delta^*_j} & \gamma_q > a_j \end{cases} \tag{2}$$

where $\Delta^*_{j-1} = \frac{a_j - a_{j-1}}{2}$ and $\Delta^*_j = \frac{a_{j+1} - a_j}{2}$ represent the boundaries of the beat length segment around a_j. Combining all annotation windows provides a normalised error sequence $\zeta_{\gamma|a}$ representing the error of the beats given the annotations.
In addition to the over-detection case, there is a similar issue related to under-detection. If very few beats are compared to a sequence of annotations, then in many instances there will no beats γ_q within each annotation window. To assign a cost to under-detection we adopt a two-way mismatch (TWM) procedure to extracting the beat error. Just as we derived $\zeta_{\gamma|a}$, we now repeat the process and compare the annotations to the beats to obtain $\zeta_{a|\gamma}$. The case of under-detection of beats to annotations is now transformed to over detection of annotations to beats. We describe how to contend with two error sequences in the following section.

2.2. Beat error histogram

An intuitive approach to extracting a measure of beat tracking accuracy from a beat error sequence ζ would be to find the variance of the data. For accurate beat tracking cases we should expect low variance, with the converse true for inaccurate, (uniformly) distributed beats. However, should the beats occur at twice the annotated metrical level, then $\zeta_{\gamma|a}$, will have approximately 50% of errors close to zero, with the remaining 50% split between errors near to 0.5 and -0.5 of a beat. For a multi-modal distribution of this type, the variance will not adequately reflect how this is a perceptually acceptable form of tapping in time to music.
A more meaningful way to extract performance from a histogram of beat error, is to look for a measure of *peakiness*, where the peaks in the histogram represent some implicit dependence between beats and annotations. To this aim we find the entropy of the error distribution, under the condition that the bin heights, x_k, for a K bin histogram, sum to unity,

$$H = -\sum_{k=1}^{K} x_k \log(x_k) \qquad \sum_{k=1}^{K} x_k = 1. \tag{3}$$

The entropy, H, will be bounded between 0 for the delta case and $\log(K)$ for the uniform case. Because we extract H directly from the bin heights of the histogram, we must take care with the number of bins used. Too few, and we will over-estimate H; too many and we will under-estimate it. To permit the possible observation of a uniform distribution (our defined worst case), for an N-length sequence of annotations we require $K \leq N$;
To select between the two error sequences $\zeta_{\gamma|a}$ and $\zeta_{a|\gamma}$, we find the sequence with the highest entropy (i.e. worst performance),

$$H_{\max} = \max(H_{\gamma|a}, H_{a|\gamma}) \tag{4}$$

where $H_{z|y}$ is the entropy of z given y.

2.3. Beat accuracy

The extracted entropy $H_{\max}$, while sufficient to distinguish between good and bad tracking performance, is not bounded over the 0-100% scale, which makes comparison with existing evaluation methods difficult. However, for any $H_{\max}$ there will be an equivalent entropy rectangular histogram (EERH), with p bins y_p of height $1/p$. By defining the total number of bins P = 100, and allowing p to take non-integer values, we can transform the entropy on to the desired 0-100% scale.
Using eqn. (3) we can show that an EERH of width p will have entropy $\log(p)$. To equate $H_{\max}$ to $\log(p)$, under the condition $K \neq P$, we scale $H_{\max}$ to be bounded between 0 and $\log(P)$,

$$\log(p) = H_{\max} - \log(K) + \log(P). \tag{5}$$

Rearranging for p, we find the more delta-like the EERH, the greater the beat accuracy, such that

$$\text{acc} = (1 - \frac{p}{P}) * 100\%. \tag{6}$$

In beat tracking evaluation, the aim is often to define beat accuracy over a large test database, e.g. [6, 8]. For each file in a test database we can define overall beat accuracy as the mean of the individual accuracies of each file, which we label μ_H. Because the beat error

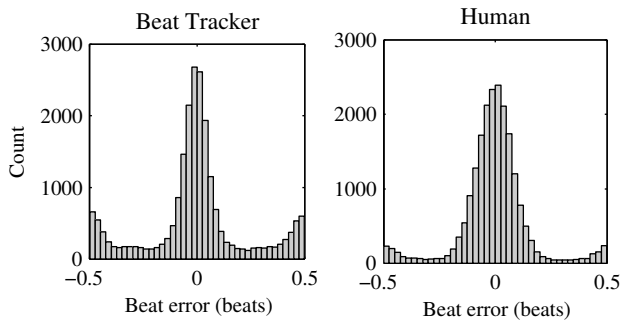

Fig. 2. Beat error histograms: (left) beat tracker and (right) human tapping performance.

Method	Beat tracker	Human tapper
CML	**54.8**	52.8
AML	78.8	**87.7**
Dix	61.5	**77.2**
Cem	55.5	**61.4**
μ_H	41.5	**53.6**
g_H	68.5	**69.7**

Table 1. Comparison of results for beat tracker and human tapping performance. Evaluation methods are CML: continuity at correct metrical level, AML: no continuity at allowed metrical levels, Dix: Dixon approach, Cem: Cemgil et al approach, g_H: Accuracy over global error histogram, μ_H: Mean accuracy over individual histograms. All accuracy values are in %.

for all files will be normalised within the [-0.5,0.5] range, we can form a single, *global* beat error histogram, which is equivalent to the bin-wise sum of the individual beat error histograms, such that the global histogram bins $X_k = \sum_{m=1}^{M} x_{k,m}$, where $x_{k,m}$ is the k^{th} bin for the m^{th} file. Given this global histogram, we can find the global entropy $H_{\max}$ and global EERH to define a second measure of beat accuracy, g_H. The principal difference between μ_H and g_H is that for μ_H all files are considered equal, whereas for g_H all beats are of equal importance.

3. COMPARISON OF EVALUATION METHODS

To investigate the properties of our proposed entropy based metric, we evaluate two beat tracking methods: our non-causal beat tracker [8] and a human tapping performance, over a 222 excerpt beat annotated database containing approximately 20,000 beats [4]. For further details on the test database and beat tracker see [8].

Before extracting any quantitative data, we can inspect the beat error histograms to infer information about the beat tracking performance. Fig. 2 shows the global beat error histograms for the beat tracking system and the human tapper. The most immediate observation is that the histograms have different shapes. The beat tracker error histogram has a tight central peak (at zero error) with two further peaks at $\pm$ 0.5 beats. The histogram of the human tapper on the other hand, has a wider central peak with much lower outer peaks. A plausible explanation for the different histogram shapes would be that the human tapper was proficient in finding the annotated metrical level (i.e. tapped at the correct tempo), but that the taps were poorly localised to the annotations. The beat tracker could be considered less consistent in metrical level selection but with more accurate localisation within the specified level. The beat tracker histogram also has a higher "noise floor", suggesting a greater proportion of cases with uniform-like error distributions, where the beats were poorly tracked.

3.1. Overall Performance Comparison

To compare evaluation methods we present results obtained from four existing evaluation methods and our approach over the same data, these are: i) *CML* - the continuity based method from [6, 8] which extracts the ratio of the longest continuously correct beat sequence to the length of the input, where beats must occur within a $\pm$ 17.5% allowance window and be at the correct metrical level; ii) *AML* - which uses an identical threshold, but defines accuracy as the ratio of the sum of the lengths of continuous segments to the length of the input, and additionally allows for tapping on the off-beat and at twice or half the annotated metrical level; iii) *Dix* - Dixon's method [1] finds the ratio of 'hits' to 'hits' plus 'false positives' plus 'false negatives', where 'hits' occur with 70ms of each annotation, 'false negatives' are the number of unmatched annotations and 'false positives' are the number of beats outside allowance windows; iv) *Cem* - the method of Cemgil et al [7] measures beat accuracy using the time between each annotation and the nearest estimated beat location on a Gaussian error function with 40ms standard deviation. To contend with false positives, the sum of the error across all annotations is divided by the mean of the number of annotations and beats. Beats far from annotations (false negatives), will be assigned an accuracy of 0% from the Gaussian error function. Results for each approach and our two entropy methods are shown in Table 1.

We can make several observations from the results presented. First, the accuracy values are quite widely spread over the 0-100% range, and furthermore that the relative ordering between the beat tracker and human is not consistent either (e.g. CML vs AML). There is also a significant difference between g_H and μ_H, which is more pronounced for the beat tracker than for the human tapping data. From this we infer that the human tapping global histogram is a truer reflection of the individual error histograms, but that the individual error histograms for the beat tracker are more varied in shape and contribute to a more uniform (higher entropy) global histogram.

3.2. Distribution of performance by file

To further investigate the differences between the existing evaluation methods, histograms of beat accuracy across the test database are shown for each method in fig. 3. Here the differences between the methods becomes more clear. *CML* is close to making a binary decision between 0 and 100%, which is largely determined by the metrical level at which the tapping occurs. *AML* confirms that many of the 0% cases from *CML* are within the allowance threshold but merely at a different metrical level. For *CML* and *AML* once all the beats occur within the defined allowance window there is no further means to distinguish performance. The *Dix* and *Cem* methods are more evenly spread over the 0-100% range, but analysis is limited in both cases to the annotated metrical level and under the condition that the beats are in phase (i.e. not tapped on the off-beat). All the off-beat cases are assigned 0% accuracy.

An interesting feature of our entropy based approach is that no cases are assigned 0% accuracy nor any with 100%. Using even relatively few histogram bins, in this case 40, to obtain 100% would require all

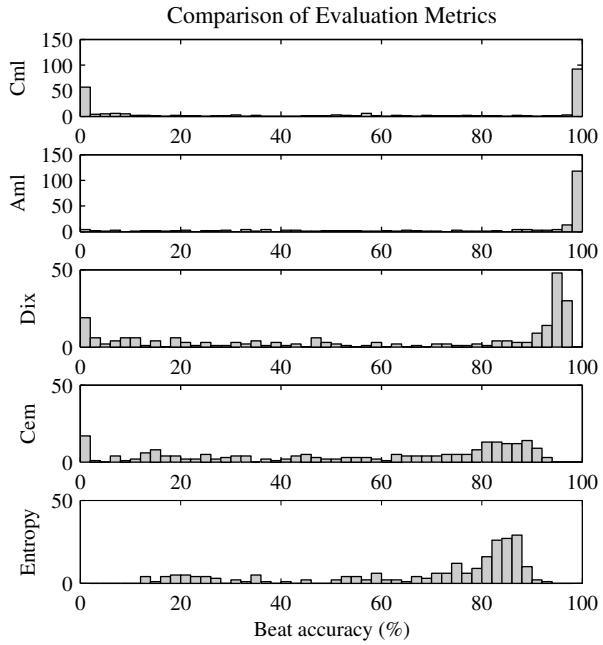

Fig. 3. Histograms of beat accuracy scores for the beat tracker over each of the 222 excerpts from the annotated database [4]. CML: continuity required at correct metrical level, AML: continuity not required at allowed metrical levels, Dix: Dixon approach, Cem Cemgil et al approach, Entropy method. Note that the vertical scales for CML and AML differ from Dix, Cem and the (μ_H) Entropy method.

beats to be within $\pm$ 1.25% of the annotations, which for a piece at approximately 120 bpm, (with beat period 500ms) would be equivalent to a 12.5ms window. This is considerably tighter than the $\pm$ 17.5% allowance window used in *CML* and *AML*, which under similar conditions would be 135ms wide. The narrow histogram bins therefore allow for the distinction between what for threshold based approaches, are indistinguishable 100% cases.

Just as 100% is an unrealistic outcome for our approach, the same is true of 0% accuracy, i.e. a perfectly uniform beat error histogram. For the other approaches, 0% can occur when beats consistently fall outside of the allowance windows, or, for all except *AML*, when beats are tapped on the off-beat. Due to the shift-invariant nature of the entropy calculation, (eqn. (3)), the off-beat, or indeed any shift, will be considered just as accurate as the on-beat. However, this only applies to the μ_H value. Since the g_H accuracy is derived from a histogram of beat error over many musical files, different locations of significant peaks will, when averaged, lead to a flatter global histogram with higher entropy. Similarly, tapping at a different metrical level to that of the annotations will lead to more peaks in the beat error histogram, and therefore lower overall accuracy.

A particular special case, known as *tempo drift*, highlights one further property of our method, with which threshold based approaches can only partially contend. It arises when the beats are tapped at a slightly incorrect tempo, and continually drift between the on and off-beat. For threshold based approaches, these beats often fall within the allowance windows and are then considered correct, even though this is regarded as a perceptually disturbing error [3]. In our method, the drifting of beats is represented by a uniform-like beat error distribution, with high entropy and therefore low beat accuracy.

The underlying feature of our entropy based approach is that it rewards a consistent relationship between beats and annotations, at a close metrical level, rather than the explicit proximity to ground truth locations as with threshold based methods.

4. CONCLUSIONS

We have explored a new approach for the evaluation of beat tracking systems, which measures beat accuracy in terms of the entropy of a beat error histogram. Our method is able to analyse all metrical levels simultaneously and is able overcome some of the limitations of existing threshold based approaches, including a rejection of the perceptually disturbing tempo drift case.

The intuitive justification for the use of thresholds in beat evaluation is that beats falling within defined allowance windows are judged to be perceptually in time, where as those outside are not. While it is possible to manually construct counter-examples to break threshold based evaluation methods, similar examples can be created to give very high accuracy for our approach, where the beats might not have any relevance to the musical input (e.g. at a precise offset of -20% of a beat). It is important to recognise that beats are perceived events, which may or may not correspond to any actual event in the audio. Therefore in the evaluation of beat tracking systems, we should aim for a metric which matches human judgement. As part of our future work we intend to undertake listening tests to examine the perceptual validity of the existing evaluation methods.

5. REFERENCES

[1] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, pp. 39–58, 2001.

[2] 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*, Evanston, IL, USA, 2004, pp. 558–562.

[3] R. B. Dannenberg, "Towards automated holistic beat tracking, music analysis, and understanding," in *Proceedings of 6th International Conference on Music Information Retrieval*, London, United Kingdom, 2005, pp. 366 –373.

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

[5] 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*, 1997, pp. 9–16.

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

[7] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering," *Journal Of New Music Research*, vol. 29, no. 4, pp. 259–273, 2001.

[8] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transcations on Audio, Speech and Language Processing*, 2007, to appear.