

ENTROPY BASED BEAT TRACKING EVALUATION

Matthew E. P. Davies and Mark D. Plumbley
Centre for Digital Music
Queen Mary, University of London

ABSTRACT

In this paper, we present a novel approach to beat tracking evaluation, based on finding the error between automatically generated beat locations and ground truth annotations. The error is normalised to the current inter-annotation-interval, such that the greatest observable error can be $\pm 50\%$ of a beat. We form a histogram of normalised beat error, from which we estimate the entropy as a measure of beat tracking performance, where low entropy indicates accurate beat locations, with the converse true for high entropy. We evaluate the performance of a human tapper in conjunction with three published beat tracking algorithms over an annotated test database and compare the results of our entropy based approach to existing evaluation methods.

Keywords – Beat tracking, evaluation, entropy

1. INTRODUCTION

The task of beat tracking of musical audio is well known and conceptually quite simple; the aim being to replicate the human ability of foot-tapping in time to a piece of music [1]. Recently, as techniques have improved, automatically extracted beat times have increasingly been used as a musically meaningful temporal segmentation for higher level analysis, for example in chord estimation [2], structural segmentation and thumb-nailing [3] and rhythmic pattern classification [4].

While many approaches exist for beat tracking (e.g. [1, 5, 6]) an important related area which has received less attention is that of evaluation – the problem of defining a suitable measure of beat accuracy. Given an agreed means of evaluation and publicly available annotated test databases (containing hand labelled beat locations), the ability to reliably compare the performance of different algorithms becomes much simpler. However, at present, there are no freely available databases, nor an agreed methodology for beat tracking evaluation. As a result, researchers use private databases [7, 5] and evaluate beat accuracy using their own metrics (e.g. [1, 5, 8, 9]), making comparative studies much harder to undertake.

For a given piece of music, where generated beat locations are compared to ground truth beat annotations, most existing methods (e.g. [1, 5, 9]) classify individual beats as correct or incorrect based on whether they occur within a pre-defined allowance window around a particular ground truth annotation. The overall accuracy for the piece is then given as the mean of the accuracy of each beat. Similarly, the mean of each piece is taken to indicate the accuracy over a test database of many pieces.

Klapuri et al [5] adopt a continuity based approach, where a given beat is only accurate if it falls within a specified allowance window and the same is true of the previous beat. Accuracy is calculated in two ways, first as the ratio of the longest continuously correct segment to the length of the input and then as the

total number of correct beats. To reflect the metrical ambiguity of human subjects tapping in time to music, both measures are re-evaluated to allow for beats to occur at double or half the annotated metrical level. This leads to multiple measures of beat accuracy, which can be harder to interpret than a single all-encompassing value of accuracy.

The difficulties with beat evaluation are not limited to generating multiple measures of accuracy. The use of thresholds in defining allowance windows around ground truth annotations can also be problematic. Dixon [1] defines a fixed window of ± 70 ms around each annotation, in which a beat must fall for it to be accurate. For songs with a fast tempo the ± 70 ms window will cover a greater proportion of the current inter-annotation-interval (IAI), (the time between successive beat annotations) than for slower songs, and will therefore be biased towards music with a faster tempo. This bias can be removed by defining a tempo-dependent threshold, where beats are accurate if they fall within 17.5% of the current IAI, as in [5, 6, 9]. However this does not address a more inherent problem of using thresholds, that a beat at the edge of an allowance window would be deemed accurate, but the same beat would be inaccurate if it occurred 1ms later, even though human listeners are incapable of discriminating between events at this time scale [10, p.29].

The motivation for our approach to beat tracking evaluation is to provide a single measure of beat accuracy that is easy to interpret and does not rely on a fixed threshold. We achieve this, not by classifying individual beat accuracy, but by measuring the *beat error*. We extract the time between each beat and the nearest ground truth annotation and normalise it by the current IAI (to give a maximum error of $\pm 50\%$ of a beat). We then formulate a histogram of normalised beat error and estimate its entropy. High entropy will result from a uniform distribution, where beats are randomly distributed across the musical piece and provide no information about the locations of the ground truth annotations. Conversely, an entropy of zero will arise when all beats are exactly equal to all annotations. To permit comparisons against other evaluation metrics, we invert and normalise the entropy to give a beat accuracy measure between 0 and 100%. We evaluate the performance of three published beat tracking algorithms and a human tapper over a large annotated test database using our entropy based measure and compare results to existing evaluation approaches.

2. APPROACH

To pursue an objective approach to beat tracking, a sequence of ground truth beat annotations is required against which the output of a beat tracker can be compared. The first step in our evaluation method is to find the error between the beats and the annotations. For each ground truth annotation a_j , we could measure the distance to the nearest beat γ_b however this would limit us to analysis of the annotated metrical level [8, 6]. To allow

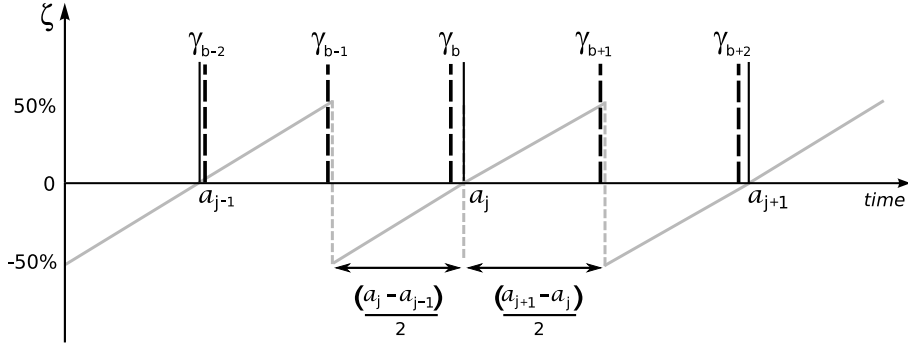


Figure 1. Calculation of beat error $\zeta_{j,q}$ from beat annotations a_j and beat locations γ_b .

us to measure each metrical level simultaneously we partition the input signal into beat length segments, each centred on a ground truth annotation, and extract all beats γ_q that fall within this range,

$$\gamma_q = \gamma_b \quad : \quad a_j - \Delta_{j-1}^* \leq \gamma_b < a_j + \Delta_j^* \quad (1)$$

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 . We now find the beat error $\zeta_{j,q}$ as the distance between each γ_q and a_j , normalised to the current IAI. The furthest any beat γ_q can be from the nearest annotation is bounded between -50% and 50% of a beat as shown in figure 1,

$$\zeta_{j,q} = \begin{cases} \frac{\gamma_q - a_j}{\Delta_{j-1}^*} & \gamma_q \leq a_j \\ \frac{\gamma_q - a_j}{\Delta_j^*} & \gamma_q > a_j \\ U & \text{no beats } \gamma_q. \end{cases} \quad (2)$$

If beats are tapped at a metrical level above the annotated locations, e.g. every other beat is close to an annotation, then there will be instances where no beats γ_q occur within each annotation centred beat segment. To overcome this missing data, we assign the error value $\zeta_{j,q} = U$, where U is a uniformly distributed random variable in the range $(-0.5, 0.5)$, equivalent to a beat error between -50% and 50% .

To provide a visual insight into the performance of a beat tracking system, we can formulate a histogram of beat error. If we believe that the beats are accurate (i.e. close to the annotations) then we can expect to observe a histogram that resembles a delta function, with a strong peak at 0% error. In the worst case, where beats are randomly distributed, and bear no relevance to the annotations, we can expect a wide, uniform-like distribution.

Figure 2 shows four example beat error histograms, those of a human tapper and three beat tracking algorithms, generated from the test database used in section 3. In addition to a strong peak centred at 0% error, we can also observe significant peaks at $\pm 50\%$ of a beat. These outer peaks occur either as result of tapping consistently on the off-beat (in anti-phase to the annotations) or tapping at twice the annotated rate, where half of the beats will be close to 0% error and the remaining beats split between $+50\%$ and -50% error.

When looking to extract a quantitative measure of beat accuracy from a beat error histogram, we might first consider the variance, expecting it to be inversely proportional to beat accuracy. However the outer peaks of the distribution, which must be considered at least partially correct, would distort the variance, where as erroneous beats that were closer to 0% error would not. As an alternative we extract the information theoretic measure of entropy [11] from the beat error histogram. In contrast to the

variance, the entropy gives a measure of the *peakiness* of the distribution, therefore rewarding beats that are consistently related to the ground truth annotations, but is blind to the metrical level or phase shift at which the beats occur. We calculate the entropy as

$$H = \sum_{k=1}^K x_k \log \frac{1}{x_k} \quad (3)$$

where there are K bins and x_k are the heights of the bins. The histogram bins are normalised such that $\sum_{k=1}^K x_k = 1$ and to maintain a real-valued output $x_k > 0$ for all k . As shown in [11] the entropy is non-negative and bounded between 0 in the best case, where beats exactly equal the annotations and $\log(K)$ for the uniform case. We can invert and normalise the entropy value to give a measure of beat accuracy \hat{H} between 0 and 100% ,

$$\hat{H} = \frac{H - \log(K)}{\log(\frac{1}{K})} * 100\%. \quad (4)$$

The normalisation removes any dependency on the number of bins used in the beat error histogram, however we empirically set $K = 40$. To give an overall measure of accuracy for a test database of multiple files we have two options. First, we can calculate the normalised entropy \hat{H}_n for each file n in the database, and find the *Mean entropy* \bar{H}_n as

$$\bar{H}_n = \frac{1}{N} \sum_{n=1}^N \hat{H}_n. \quad (5)$$

Alternatively, we can form a single histogram of beat error for all files in the database (as shown in figure 2) with the entropy calculated once using eqn. (3) but replacing x_k with the mean bin height \bar{x}_k over all N files,

$$\bar{x}_k = \frac{1}{N} \sum_{n=1}^N x_{n,k}. \quad (6)$$

After which we can normalise the entropy using eqn. (4) to give a measure of *Global entropy* $\hat{H}_{\bar{n}}$.

3. RESULTS

We include results for our evaluation measures over a beat annotated test database containing 222 files, each one minute in length over six musical genres: Dance, Rock, Jazz, Classical, Folk and Choral. For further details see [7, 6]. For comparison against our entropy based approaches, we include the evaluation method used by Dixon [1] as well as the continuity based method of Klapuri et al [5] also used in [6].

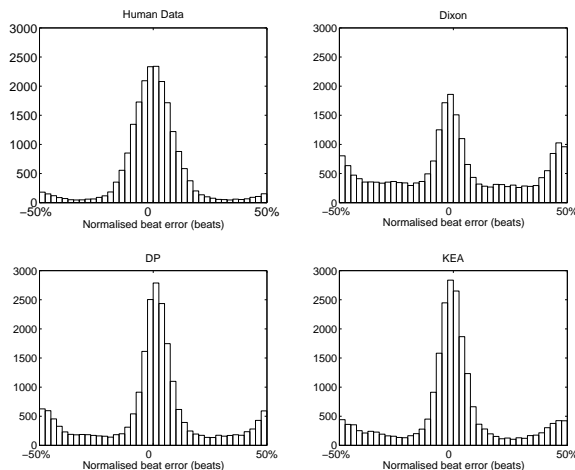


Figure 2. Beat error histograms. Clockwise from Top Right: Dixon [1], Klapuri et al [5] labelled KEA, Davies and Plumbley [6] labelled DP, and a Human tapping performance.

Dixon’s [1] measure of beat accuracy is calculated as follows

$$\text{Dixon}_{\text{acc}} = \frac{\text{hits}}{\text{hits} + F^+ + F^-} * 100\% \quad (7)$$

where a ‘hit’ is defined as a beat that occurs within $\pm 70\text{ms}$ of an annotated beat location. F^+ refers to the number of false positives, i.e. the number of beats which are not matched to any annotation and F^- which refers the number of unmatched annotations, or false negatives.

The continuity based beat evaluation approach [5, 6] differs from Dixon’s [1] method as it requires beats to be *continuously* correct. Beats must be within 17.5% of the nearest annotation *and* the local tempo must not differ by more than 17.5%. In all, four measures of beat accuracy can be calculated. These allow for continuous tracking at the correct metrical level (CML cont), the total number of correct beats at the correct metrical level (CML total). Both measures can then be recalculated allowing for the beats to be tapped at twice or half the annotated level, referred to as the allowed metrical levels, giving AML cont. and AML total. We retain only the strictest, CML cont. and the least strict, AML total. A more detailed description may be found in [6].

Results for our entropy based approach and those described above are given in Table 1 illustrating the performance of a human tapper and three published beat tracking algorithms: Dixon [1], Davies and Plumbley [6] (labelled DP) and Klapuri et al [5] (labelled KEA).

4. DISCUSSION

The first observation about the results is that the entropy based measures appear much stricter than the other evaluation methods. The principal reason for this being that 100% accuracy is not possible unless the beats are equal to the annotations, where as with the Dixon [1] and continuity based approaches [6] perfect tracking is possible if beats are consistently within the allowance windows. We should also note that the relative ordering the algorithms is not consistent across all evaluation measures. The Mean entropy and CML cont. place the human tapper as less accurate than KEA and DP, although for the remaining measures the human is most accurate. Under all measures the Dixon [1] approach is weakest. Intuitively we should expect the human

tapper to be the highest performing, and as shown in [6] the human is more successful in finding the correct metrical level and the on-beat than the algorithmic approaches, but that beat localisation is often poorer. We can confirm this by inspection of the beat error histograms in figure 2. Very few beats occur at $\pm 50\%$ for the human but the main peak, centred on 0% error, is wider than that of KEA and DP. In order for the Mean entropy for the human to be lower than KEA and DP, we infer that a greater proportion of very accurate cases with high normalised entropy \hat{H}_n , outweigh the more consistent, but generally less accurate, human performance. This suggests that the human beat error histogram is a more realistic representation of the performance for a given file, but the algorithmic histograms will vary from inaccurate uniform-like up to accurate delta-like distributions. We therefore believe the Global entropy measure to be more informative than the Mean entropy.

To further inspect the differences between the evaluation measures, we analysed each evaluation method with artificial input data. To create the artificial data we took the ground truth beat annotations and gradually degraded the performance, perturbing beats with uniform distributions of increasing width. Figure 3 shows the effect of increasing the width of the uniform distribution from a relative error of 2.5% of a beat, in 2.5% steps, up to totally random with 100% beat error. We can observe that all but the entropy approaches remain 100% accurate even for a beats perturbed by a uniform distribution of width of 20% (i.e. $\pm 10\%$ around each annotation). We also note that none of the approaches are linear. Therefore observing an accuracy of 80% for algorithm A would not be twice as good as algorithm B at 40%, as the numbers alone might suggest. It can be shown that the entropy of an exactly uniform distribution with a fixed number of bins is equal to $\log(n)$, where there are n bins of height $1/n$. Although the Global entropy curve is not linear, it is at least log-linear, so some meaningful relative comparison can be made between competing algorithms. The Mean entropy curve is always greater than or equal to the Global entropy curve, as the two can only be equal when every beat error distribution is perfectly uniform, as any irregularities in bin height will reduce the entropy and therefore increase the normalised beat accuracy. A current limitation of our approach is that the entropy calculation is *shift-invariant* with respect to the ordering of the his-

Beat Tracker	Mean Entropy (%)	Global Entropy (%)	CML Cont. (%)	AML Total (%)	Dixon Acc. (%)
Human	33.7	20.1	52.8	87.7	77.2
DP[6]	36.8	14.2	54.8	78.7	61.5
KEA[5]	38.3	15.5	55.8	80.1	64.6
Dixon[1]	23.2	5.5	21.9	52.0	35.4

Table 1. Beat accuracy results. The performance of a human tapper with three published algorithms are compared over our two proposed entropy measures, two continuity based measures from Davies and Plumbley [6] and Dixon’s [1] approach.

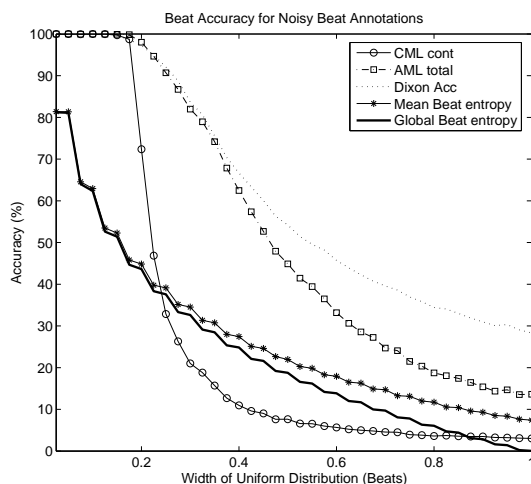


Figure 3. Comparison of beat tracking accuracy for beat annotations perturbed by a white noise process.

togram bins. Although unlikely, beats consistently 10% ahead of the beat would be considered just as accurate to beats centred on 0%, even though the beats themselves would be perceptually out of time with the input. We intend to investigate this further, possibly incorporating the mean of the distribution into the beat accuracy calculation.

5. CONCLUSIONS

We have presented an entropy based method for the evaluation of beat tracking systems. We extract the entropy from a distribution of normalised beat error, and scale the calculated value to indicate beat accuracy between 0 and 100%. We have demonstrated that the beat error histogram is a useful visualisation of beat tracking performance, within which the simultaneous analysis of multiple metrical levels is possible, a feature not present in other approaches to beat tracking evaluation. Although beat accuracy from our entropy measure suggests poorer performance than published approaches, it is an analytically meaningful statistic which is not reliant on pre-defined parametric thresholds. As part of our future work we intend to further investigate the validity of using entropy to evaluate beat trackers, and plan to conduct listening tests to infer which evaluation methods are most perceptually meaningful.

6. ACKNOWLEDGEMENTS

MEPD is supported by a college studentship from Queen Mary University of London. This research has been partially funded by EPSRC grants GR/S75802/01 and GR/S82213/01.

7. 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] J. P. Bello and J. Pickens, “A robust mid-level representation for harmonic content in music signals,” in *Proceedings of 6th International Conference on Music Information Retrieval*, London, United Kingdom, 2005, pp. 304 – 311.
- [3] M. Levy, M. Sandler, and M. Casey, “Extraction of high level musical structure from audio data and its application to thumbnail generation,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.
- [4] S. Dixon, F. Gouyon, and G. Widmer, “Towards characterisation of music via rhythmic patterns,” in *Proceedings of 5th International Conference on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 509–517.
- [5] 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.
- [6] M. E. P. Davies and M. D. Plumbley, “Context-dependent beat tracking of musical audio,” Tech. Rep., Queen Mary University of London, Centre for Digital Music, 2006, <http://www.elec.qmul.ac.uk/people/markp/2006/C4DM-TR-06-02.pdf>.
- [7] S. Hainsworth, *Techniques for the Automated Analysis of Musical Audio*, Ph.D. thesis, Department of Engineering, Cambridge University, 2004.
- [8] 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, 2000.
- [9] 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.
- [10] J. London, *Hearing in Time: Psychological Aspects of Musical Meter*, Oxford University Press, 2004.
- [11] M. D. Plumbley, “On information theory and unsupervised neural networks,” Tech. Rep., Cambridge University, Department of Engineering, 1991.