Multi-Feature Beat Tracking

ARTICLE *in* IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING · APRIL 2014

DOI: 10.1109/TASLP.2014.2305252

3 AUTHORS:

Jose R Zapata
Universidad Pontificia Bolivariana
10 PUBLICATIONS **55** CITATIONS
SEE PROFILE

Matthew Davies
Institute for Systems and Computer Engine…
52 PUBLICATIONS **453** CITATIONS
SEE PROFILE

Emilia Gómez
University Pompeu Fabra
91 PUBLICATIONS **1,123** CITATIONS
SEE PROFILE

Multi-Feature Beat Tracking

José R. Zapata, Matthew E. P. Davies, and Emilia Gómez

***Abstract*—A recent trend in the field of beat tracking for musical audio signals has been to explore techniques for measuring the level of agreement and disagreement between a committee of beat tracking algorithms. By using beat tracking evaluation methods to compare all pairwise combinations of beat tracker outputs, it has been shown that selecting the beat tracker which most agrees with the remainder of the committee, on a song-by-song basis, leads to improved performance which surpasses the accuracy of any individual beat tracker used on its own. In this paper we extend this idea towards presenting a single, standalone beat tracking solution which can exploit the benefit of mutual agreement without the need to run multiple separate beat tracking algorithms. In contrast to existing work, we re-cast the problem as one of selecting between the beat outputs resulting from a single beat tracking model with multiple, diverse input features. Through extended evaluation on a large annotated database, we show that our multi-feature beat tracker can outperform the state of the art, and thereby demonstrate that there is sufficient diversity in input features for beat tracking, without the need for multiple tracking models.**

***Index Terms*—Beat tracking, evaluation, music information retrieval, music signal processing.**

I. Introduction

THE extraction of beat times from musical audio signals is a key aspect of computational rhythm description [1], and forms an important research topic within music information retrieval (MIR). Since the earliest audio beat tracking systems [2]–[4] in the mid to late 1990s, there has been a steady growth in the variety of approaches developed and the applications to which these beat tracking systems have been applied. For a recent review see [5, ch.2]. Indeed, beat tracking systems are now considered "standard" processing components within many MIR applications, such as chord detection [6], structural segmentation [7], cover song detection [8], automatic remixing [9] and interactive music systems [10].

While the efficacy of beat tracking systems can be evaluated in terms of their success of these end-applications, e.g. by measuring chord detection accuracy, considerable effort has been placed on the evaluation of the beat tracking systems directly through the use of annotated test databases in particular within the MIREX initiative. In the small number of comparative studies of automatic beat tracking algorithms with human tappers [11]–[15] musically trained individuals are generally shown to be more adept at tapping the beat than the best computational systems. Given this gap between human performance and computational beat trackers, we consider that beat tracking is not yet a solved problem and still has high potential for improvement. In recent work [14], it was speculated that the advancement of computational beat tracking systems was stagnating due to a lack of diversity in annotated datasets, and the pursuit of so-called "universal" models for beat tracking which attempt to use a single approach to determine beat locations in all styles of music. Genre-specific breakdowns of beat tracking performance (e.g.[12], [13], [16]) illustrate a heavy preference towards what could be considered "easier" styles of music, such as rock, pop, and electronic dance music, which also tend to be the most abundant within current annotated datasets.

While the idea of a universal model for beat tracking would seem to be an attractive goal, Collins [15] proposes strong arguments as to why this is (currently) unrealistic. He suggests that the main flaw of computational beat tracking systems is a lack of understanding of the higher-level musical context; however, this context is obvious to the trained human listener when tapping to music. The eventual route towards improving beat tracking would therefore appear to be through the use of higher level knowledge of musical style coupled with the understanding of how to apply this knowledge in the context of beat tracking. Through simulated evaluation (e.g., in [17], [18, ch.4]), where *a priori* knowledge of the best beat tracking system per genre can be used, large hypothetical gains in performance are possible. However, to the best of our knowledge, no such system currently exists which can outperform the state of the art using automatic determination of musical style or genre.

Where improvements to the state of the art have been made, is through a more indirect usage of the effectiveness of different beat tracking systems for different types of music signals. In [14], a measure of mutual agreement (*MA*) was used to select between a committee of five existing state of the art beat tracking algorithms, where the beat tracking output which agreed most with the remainder of the committee was chosen

Manuscript received August 13, 2013; revised November 26, 2013; accepted January 30, 2014. Date of publication February 07, 2014; date of current version February 19, 2014. This work was supported in part by the R+I+D Ph.D. scholarship of Universidad Pontificia Bolivariana and Colciencias (Colombia), the projects of the Spanish Ministry of Science and Innovation DRIMS (MICINN-TIN2009-14247-C02-01), SIGMUS (MINECO-TIN2012-36650) and Mires (EC-7PM-MIReS), and in part 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) by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT), the FCT post-doctoral grant (SFRH/BPD/88722/2012), the PHENICX (EU: FP7 2007/2013 grant agreement n 601166) and COFLA (Junta de Andaluca P09-TIC-4840) projects. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Thushara Dheemantha Abhayapala.

J. R. Zapata is with the Faculty of TIC, Universidad Pontificia Bolivariana, Medellin, Colombia (e-mail: joser.zapata@upb.edu.co).

M. E. P. Davies is with INESC TEC, 4200-465 Porto, Portugal (e-mail: mdavies@inescporto.pt).

E. Gómez is with the Music Technology Group (MTG), Universitat Pompeu Fabra, 08018 Barcelona, Spain (e-mail: emilia.gomez@upf.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2305252

as the most representative. Although this system was shown to improve upon the overall performance of any individual algorithm on a large database, it does not represent a practical beat tracking solution since it requires the execution of five separate beat tracking algorithms across multiple platforms. Here, our motivation is to transfer this concept mutual agreement in beat tracking towards a standalone beat tracking solution.

Following recent results in [18, ch.4] which show greater potential for improving beat tracking through input features to beat tracking systems rather than tracking models, we present a mutual agreement based beat tracker which draws information from multiple input features and passes them to a single beat tracking model. Our hypothesis is that there is sufficient diversity within different types of input features to facilitate an improvement in the state of the art by selecting between the resulting beat sequences, without the need for multiple separate tracking models. Furthermore, we speculate that the improvement obtained in the multiple beat tracking system in [14] was the result of the different input features to the beat tracking systems, rather than the inherent properties of the different beat tracking models. Therefore, to best address hypothesis, we mirror the main processing steps in [14] for the calculation of mutual agreement. However, in addition to the main methodology of [14], we present a more extensive evaluation, we demonstrate how to use the mean *MA* value as a measure of beat tracking confidence, and we examine the beat tracking committees in terms of computational complexity.

The remainder of the paper is structured as follows: in Section II we describe the set of input features, summarize the beat tracking model and present the methods we use to measure agreement in beat sequences. In Section III we describe the experimental setup in terms of the test database and evaluation methods used. In Section IV we present an extended evaluation. This includes measuring the performance of each individual onset detection function and then demonstrating the improvement when selecting a beat sequence using mutual agreement. Section V concludes the paper with discussion of the results and areas for future work.

II. Multi-Feature Beat Tracking System

The proposed multi-feature beat tracking system (shown in Fig. 1) is composed of three parts, first, a set of onset detection functions (ODF), this is followed by beat period estimation and beat tracking for each ODF. Finally, the overall beat output is chosen using a selection method applied to the set of estimated beat locations. The proposed beat tracker is publicly available[1].

A. Input Features

In beat tracking, an onset detection function is commonly used as a mid-level representation that reveals the location of transients in the original audio signal. This onset detection function is designed to show local maxima at likely event locations [19]. Many methods exist to emphasize the onset of musical events and the performance of beat trackers strongly depends on the low-level signal features used at this stage [20].

[1]http://essentia.upf.edu/, *BeatTrackerMultiFeature()*, Affero-GPL.

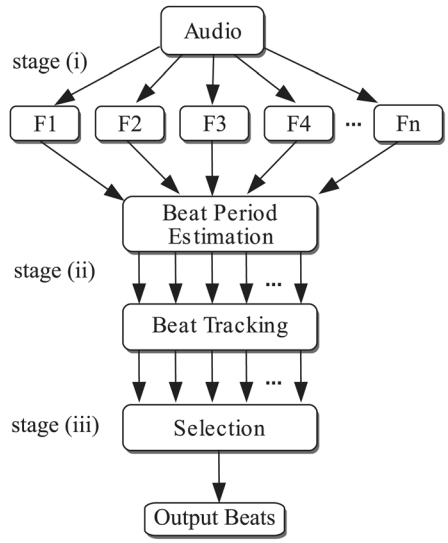

Fig. 1. System Overview. The multi-feature beat tracker is comprised of three stages: (i) a set of onset detection functions, $F1 \ldots Fn$, as input features; (ii) beat period estimation and beat tracking; and iii) a selection method to choose between the set of beat outputs.

Towards building our multi-feature beat tracking system we first collected the onset detection functions from each beat tracking algorithm used in [14] and [21]. Some of these algorithms were freely available online and the remainder were provided by the algorithm authors or reimplemented. In addition, other onset detection functions were included, where they were deemed to be complementary to those already selected, i.e. those with the ability to detect note onsets in specific musical contexts such as music without strong percussive content [22]. As in [14] our goal is to obtain a small but diverse committee making use of publicly available reference implementations wherever possible. In addition, while a computationally efficient system is not the specific goal of this research, we seek to avoid any input features which are very computationally expensive to calculate - as their eventual benefit may not be worth the increase in computation time.

In total we compiled an initial set of nine onset detection functions which are described below. Note that while each onset detection function is extracted according to its original parameterization in terms of window length and hop size (assuming a mono input audio signal sampled at 44.1 kHz), each onset detection function is subsequently resampled to have a temporal resolution of 11.6 ms prior to extracting the beats in order to match the input feature resolution expected by the beat tracking model. In the following equations for generating onset detection functions, $X(k)$ refers to the discrete Fourier transform spectrum of an audio frame x_n, the symbol k is the index over linear frequency bins in X and b is an index over a smaller number of sub-bands, B.

1) Energy Flux (EF): Equation (1) is a simplified implementation of the Energy flux function [23], and is calculated by computing short time Fourier transform frames using a window size of 2048 and hop size of 512, corresponding to an input feature resolution of 11.6 ms. From these frames, each input feature sample $EF(n)$ is calculated as the magnitude of the differences of the root mean square (RMS) value between the current short time Fourier transform frame and its predecessor:

$$EF(n) = |RMS(X_n(k)) - RMS(X_{n-1}(k))|. \quad (1)$$

2) Spectral Flux (SFX): The spectral flux onset detection function proposed in [24] and presented in (2), is calculated by computing short time Fourier transform (STFT) frames using a window size of 2048 and hop size of 512, corresponding to an input feature resolution of 11.6 ms. From these frames, each input feature sample $SFX(n)$ is calculated as the sum of the positive differences in magnitude between each frequency bin of the current short time Fourier transform frame and its predecessor:

$$SFX(n) = \sum_{k=1}^{K} H(|X_n(k)| - |X_{n-1}(k)|). \quad (2)$$

where $H(x) = \frac{x-|x|}{2}$ is the half-wave rectifier function.

3) Spectral Flux Log Filtered (SFLF): Introduced by Böck *et al.* [25] this method is based on spectral flux, but the linear magnitude spectrogram is filtered with a pseudo Constant-Q filter bank, as can be seen in (3),

$$X_n^{logfilt}(b) = log(\lambda \cdot (|X_n(k)| \cdot F(k,b)) + 1). \quad (3)$$

where the frequencies are aligned according to the frequencies of the semitones of the western music scale over the frequency range from 27.5 Hz to 16 kHz, using a fixed window length for the STFT, a window size of 2048 and a hop size of 512. The resulting filter bank, $F(k,b)$, has $\mathrm{B} = 82$ frequency bins with b denoting the bin number of the filter and k the bin number of the linear spectrogram. The filters have not been normalized, resulting in an emphasis of the higher frequencies, similar to the high frequency content (*HFC*) method. From these frames, in (4) each input feature sample is calculated as the sum of the positive differences in logarithmic magnitude (using λ as a compression parameter, $\lambda = 20$) between each frequency bin of the current STFT frame and its predecessor:

$$SFLF(n) = \sum_{b=1}^{B=82} H\left(|X_n^{logfilt}(b)| - |X_{n-1}^{logfilt}(b)|\right). \quad (4)$$

4) Complex Spectral Difference (CSD): The complex spectral difference input feature [26], presented in (5), is calculated from the short time Fourier transform of 1024 sample frames with a 512 sample hop size, resulting in a resolution of 11.6 ms. The feature produces a large value if there is a significant change in magnitude or deviation from expected phase values, different from the spectral flux that only computes magnitude changes in frequency. $\tilde{X}_n$ is the expected target amplitude and phase for the current frame and is estimated based on the values of the two previous frames assuming constant amplitude and rate of phase change,

$$CSD(n) = \sum_{k=1}^{K} |X_n(k) - \tilde{X}_n(k)|. \quad (5)$$

5) Beat Emphasis Function (BEF): Introduced in [27], the Beat emphasis function is defined as a weighted combination of sub-band complex spectral difference functions (5), $S_b(n)$, which emphasize periodic structure of the signal by deriving a weighted linear combination of 20 sub-band onset detection functions driven a measure of sub-band beat strength,

$$BEF(n) = \sum_{b=1}^{B=20} w(b) \cdot S_b(n). \quad (6)$$

where the weighting function $w(b)$ favours sub-bands with prominent periodic structure. In (6), BEF is calculated from the short time Fourier transform of 2048 sample frames with a 1024 sample hop size, the output is interpolated by a factor of two, resulting in a resolution of 11.6 ms.

6) Harmonic Feature (HF): The harmonic feature presented by Hainsworth and Macleod [28] is a harmonic change detection and is calculated in (7) by computing a short time Fourier transform using a window size of 2048 sample frames with a 512 sample hop size. *HF* uses a modified Kullback-Leibler distance measure to detect spectral changes between frequency ranges of consecutive frames. The modified measure is thus tailored to accentuate positive energy change,

$$HF(n) = \sum_{b=1}^{B} log_2\left(\frac{|X_n(b)|}{|X_{n-1}(b)|}\right). \quad (7)$$

Only the region of 40 Hz-5 kHz was considered to pick peaks, a local average of the function was formed and then the maximum picked between each of the crossings of the actual function and the average.

7) Mel Auditory Feature (MAF): The Mel Auditory Feature, introduce by [29], is calculated by resampling the audio signal to 8 kHz and calculating a short time Fourier transform magnitude spectrogram with a 32 ms window and 4 ms hop size. In (8) each frame is then converted to an approximate "auditory" representation in 40 bands on the Mel frequency scale and converted to dB, $X^{mel}(b)$. Then the first order difference in time is taken and the result is half-wave rectified. The result is summed across frequency bands before some smoothing is performed to create the final feature,

$$MAF(n) = \sum_{b=1}^{B=40} H\left(|X_n^{mel}(b)| - |X_{n-1}^{mel}(b)|\right). \quad (8)$$

8) Phase Slope Function (PSF): The group delay is used to determine instants of significant excitation in audio signals and is computed as the derivative of phase over frequency $\tau(k)$, as seen in (9). In [22], it was used as an onset detection function. Using an analysis window with a large overlap the average group delay was computed for each window. The obtained sequence of average group delays is referred to as the phase slope function (PSF). The resulting resolution of the signal is 6.2 ms. To avoid the problems of unwrapping the phase spectrum of the signal for the computation of group delay can be computed as:

$$\tau(k) = \frac{X_{\mathbb{R}e}(k) \cdot Y_{\mathbb{R}e}(k) + X_{\mathbb{I}m}(k) \cdot Y_{\mathbb{I}m}(k)}{|X(k)|^2}. \quad (9)$$

Where $X(k)$ and $Y(k)$ are the Fourier Transforms of $x[n]$ and $nx[n]$, respectively. The phase slope function is then computed as the negative of the average of the group delay function.

9) Bandwise Accent Signals (BAS): Introduced by Klapuri *et al.* [16], Bandwise Accent Signals are calculated from 1024 sample frames with a 512 sample hop size. The Fourier transform of these frames is taken and used to calculate power envelopes at 36 sub-bands on a critical- band scale. Each sub-band is up-sampled by a factor of two, smoothed using a low-pass filter with a 10-Hz cutoff frequency and half-wave rectified. A weighted average of each band and its first order differential is taken, $u_b(n)$. In [16] each group of 9 adjacent bands (i.e. bands

1–9, 10–18, 19–27 and 28–36) are summed up to create a four channel (c) input feature with a resolution of 5.8 ms, however in this paper we simply sum all 36 power envelopes to generate a single output feature,

$$BAS(n) = \sum_{b=1}^{36} u_b(n). \tag{10}$$

B. Beat Period Estimation and Tracking Model

Given each onset detection function we now address the task of estimating beat locations. Since our system relies on a single beat tracking model, we select a beat tracker which has been shown to perform well in comparative studies and is freely available. To this end, we choose the method of Degara *et al.* [30], which was also part of the committee of beat tracking algorithms in [14].

The core of Degara's beat tracking model is a probabilistic framework which takes as input an onset detection function (used to determine the phase of the beat locations) and a periodicity path which indicates the predominant beat period (or tempo) through time. While a different input feature (or user-specified input) could be used to determine the periodicity path, in practice it is estimated from the same onset detection function. The technique for finding the periodicity (as used in [30]) is an offline version of the Viterbi model in [31]. This Viterbi model assumes the beat period to be a slowly varying process with transition probabilities modeled using a Gaussian distribution of fixed standard deviation. To estimate the beats, the system integrates musical-knowledge and signal observations using a probabilistic framework to model the time between consecutive beat events and exploits both beat and non-beat signal observations. For more information on the tracking method, see [30]. Since our primary concern in this paper relates to the input features supplied to the beat tracker, we can treat the beat tracker as a "black box".

To create the committee of beat trackers, we calculate a separate periodicity path and set of beat locations for each onset detection function.

C. Selection Method and Measuring Mutual Agreement

The mutual agreement (*MA*) method was presented in [14] and [21] to compare multiple beat tracking sequences. When looking to select one beat sequence from the committee, the beat sequence with the maximum mutual agreement (*MaxMA*) was chosen. When evaluated, this method was shown to provide significant improvements in beat tracking performance over 16 reference beat tracking systems [14].

As shown in Fig. 2 and equation (11), the mutual agreement, MA_i, of a sample is computed by using the beat estimations of a committee of N beat trackers on a musical piece, measuring the agreement $A_{i,j}$ between estimated beat sequences i and j.

$$MA_i = \frac{1}{N-1} * \sum_{j=1, j\neq i}^{N} A_{i,j}. \tag{11}$$

The Mean Mutual Agreement (*MMA*) is computed by measuring the mean of all N mutual agreements MA_i between all estimated beat tracker outputs i.

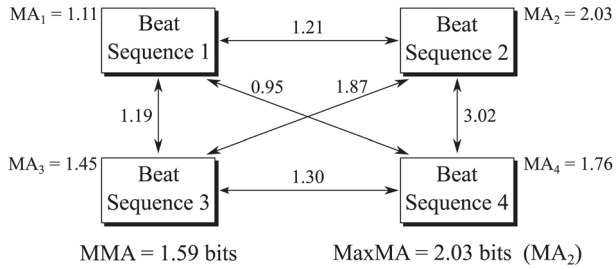

Fig. 2. Example calculation of the Mutual Agreement (MA) and Maximum Mutual Agreement (MaxMA) for a song with the beats estimated from a committee of four beat trackers.

In [32] the properties of existing beat tracking evaluation measures [33] were reviewed for the purpose of measuring mutual agreement. Of these, the Information Gain approach [34] (InfGain) was shown to have a true zero value able to match low *MMA* (measured in bits) with unrelated beat sequences. In [21] an *MMA* value of 1.5 bits was shown to work as a confidence threshold for beat tracking.

While Information Gain was shown to be a good indicator of agreement between beat sequences from among existing beat tracking evaluation methods, it is not the only approach which could be used. In this paper we also explore an alternative mechanism for measuring agreement, the *regularity function* of Marchini and Purwins [35], which quantifies the degree of temporal regularity between time events. To calculate the regularity we first concatenate and sort the beats of two different beat sequences, then we compute the histogram of the time differences between all possible combinations of two beats (the complete inter-beat interval histogram, *CIBIH*). In this way, we obtain is a kind of "harmonic series" of peaks that are more or less prominent according to the self-similarity of the sequence at different time scales. Second, we compute the autocorrelation $ac(t)$ (where t corresponds to lag in seconds) of the *CIBIH* which, in the case of a regular sequence, has peaks at multiples of the tempo. Let t_{usp} be the positive time value corresponding to its upper side peak. Given the sequence of m beats $x = (x_1, \ldots, x_m)$ we define the regularity of the sequence of beats x to be:

$$Regularity(x) = \frac{ac(t_{usp})}{\frac{1}{t_{usp}} \int_0^{t_{usp}} ac(t_{usp})}. \tag{12}$$

If the beat estimations are more equally spaced in time the regularity value will be higher, whereas if the beat estimations are unrelated the regularity value will be lower. For more information see [35].

Referring again to Fig. 1, the chosen selection mechanism (either Information Gain or Regularity) is the final stage in our multi-feature beat tracking which provides the eventual beat output.

III. Experimental Setup

A. Dataset

The largest available dataset for beat tracking evaluation to date was introduced by Gouyon [36] and contains a total

of 1360 excerpts from different styles of music, in a compilation of the Klapuri [16], Hainsworth [37], SIMAC[2] project and CUIDADO [38] project datasets. It will be referred to as *Dataset1360* throughout this paper.

We use *Dataset1360* to analyze the diversity and accuracy of the different onset detection functions. Based on these results we will: i) select our committee of onset detection functions; ii) give empirical evidence of using Maximum Mutual Agreement (*MaxMA*) for selecting the best beat tracking estimation from the committee; and iii) verify the behavior of the *MMA* method calculated with a committee formed by different onset detection functions to assess difficulty for automatic beat tracking.

The dataset is comprised of 10 genres: Acoustic (84 pieces), Jazz/Blues (194 pieces), Classical (204 pieces), Classical solo (79 pieces), Choral (21 pieces), Electronic (165 pieces), Afro-American (93 pieces), Rock/Pop (334 pieces), Balkan/Greek (144 pieces), and Samba (42 pieces). Inspection of the breakdown across musical genres reveals fewer than 40% of excerpts are Rock, Pop or Electronic which are often considered "easier" categories for beat tracking [14]. Given the high proportion of more challenging musical genres for beat tracking and the large number of annotated excerpts, we believe this forms a sufficiently diverse test set for measuring beat tracking performance and to demonstrate the potential gains available from our multi-feature approach. For further details on the dataset, see [36].

B. Evaluation Measures

For evaluating the beat tracking accuracy against manual annotations, we use a subset of methods from the beat tracking evaluation toolbox [33]. These evaluation methods are also used in the beat tracking evaluation task within MIREX.

Among all the proposed evaluation metrics, we use the continuity measures as originally defined in [16], [37] with an output range between [0 - 100]%. This allows us to analyze both the ambiguity associated with the annotated metrical level and the continuity in the beat estimates. These accuracy measures consider regions of continuously correct beat estimates relative to the length of the audio signal analyzed. Continuity is enforced by defining a tolerance window of 17.5% relative to the current inter-annotation-interval. To allow the beat tracker to initially induce the beat, events within the first five seconds of each excerpt are not evaluated. The continuity-based criteria used for performance evaluation are the following:

- **CMLc** (Correct Metrical Level with continuity required) gives information about the longest segment of continuously correct beat tracking.
- **CMLt** (Correct Metrical Level with no continuity required) accounts for the total number of correct beats at the correct metrical level.
- **AMLc** (Allowed Metrical Level with continuity required) is the same as **CMLc** but it accounts for ambiguity in the metrical level by allowing for the beats to be tapped at double or half the annotated metrical level.
- **AMLt** (Allowed Metrical Level with no continuity required) is the same as **CMLt** but it accounts for ambiguity in the metrical level.

[2]http://mtg.upf.edu/simac/

C. Reference Systems

To compare our system against existing beat trackers, we compiled a set of existing algorithms, including those with freely available implementations online and others provided by the authors of the systems on request. To summarize the accuracy of each beat tracking system, the mean value of the performance measures across all the audio files of the test database is presented. Statistically significant differences on the mean values were also checked. For this, we use a paired T-test with $\alpha = 0.05$ as a guide to indicate statistical significance. In total we compiled 18 state of the art beat trackers (expanding the set originally in [14]) and also compare against the five committee beat tracker from [14].

IV. Results

A. Committee Members

Before presenting comparative results against other beat tracking algorithms we first analyze the composition of the committee of beat trackers in our multi-feature beat tracker. The initial committee is composed of the beat tracking outputs from the following onset detection functions, as described in Section II-A: bandwise accent signal (*BAS*), beat emphasis function (*BEF*), complex spectral difference (*CSD*), energy flux (*EF*), harmonic feature (*HF*), mel auditory feature (*MAF*), phase slope function (*PSF*), spectral flux (*SFX*) and spectral flux log filtered (*SFLF*). Following [14], we aim to reduce this initial committee to a smaller subset by including only those input features which, in combination, can lead to a hypothetical improvement in performance against the ground truth; that is, if the beats from a particular input feature are never more accurate than beats from another input feature, then there is little value in including the "weaker" input feature.

Towards determining a sub-committee, we evaluate the mean performance of each feature as input to the Degara beat tracker on *Dataset1360* in Table I. From inspection of the table, we can see the *CSD* has the highest accuracy under **AMLt**, and the *EF* and *PSF* perform least well. We speculate that this relatively low performance across the diverse set of musical styles in the test dataset is due to the specific emphasis in detecting changes in only one signal variable (i.e. energy, or phase), compared to the more general nature of the *CSD* method. Note however that overall performance itself is *not* a reliable indicator for inclusion in the sub-committee. Since we wish to exploit the ability of different input features to be appropriate for beat tracking in different contexts, our goal is towards finding a complementary set to form the committee.

To find the relevance of each ODF in the committee we make use of the sequential forward selection, SFS, method, as used in [14]. At this stage we do not make use of any methods for measuring mutual agreement–these will follow once we have selected our committee. We begin by fixing the first member of the committee as the one whose mean performance across the entire dataset is highest, in this case the *CSD* onset detection

TABLE I
MEAN CONTINUITY MEASURES PERFORMANCE (%) OF EACH FEATURE AND THE ORACLE IN THE 1360 SONG DATASET

ODF	CMLc	CMLt	AMLc	AMLt
CSD	**46.1**	**50.3**	**69.8**	**77.6**
HF	38.5	45.6	62.0	73.7
PSF	31.1	35.2	61.3	69.9
EF	39.6	44.7	56.1	64.6
SFLF	44.2	48.0	68.9	76.8
BEF	38.0	42.2	65.5	73.5
BAS	43.0	46.8	68.5	76.4
MAF	42.2	46.8	63.9	73.0
SFX	43.2	47.9	65.8	73.9
Oracle	64.9	69.0	85.5	90.5

function. To determine the second ODF to enter the committee, we proceed as follows:

(i) We choose any of the other ODFs and, for this ODF, we find the best possible beat accuracy (i.e. *oracle*) score that could be achieved by perfectly selecting between this ODF and *CSD* for every file in the dataset.

(ii) We then repeat this process to obtain a mean oracle score per potential committee member for all other ODFs across the dataset.

(iii) We then select the ODF which, in combination with the *CSD* beat tracker, leads to the maximum improvement in overall beat tracking accuracy over using just *CSD* on its own.

Once the second committee member has been added, we then remove it from the pool of potential committee members and repeat the process described above. However, rather than comparing to the beat accuracy scores from *CSD* per file in the dataset, we update this to reflect the best score per file from the two committee members. This procedure is iteratively continued until all onset detection functions have been included, with the same merging of best scores at each iteration.

When the selection process has been completed we can look at the order in which each ODF entered the committee and the improvement in performance achieved by its inclusion. We can then determine a subset by fixing the number of committee members at the point where improvements offered by additional members is small. Using the SFS method the order in which the ODF enter the committee in the order listed in Table I, i.e. *CSD*, *HF*, *PSF* etc.

In Fig. 3(a) a comparison between the mean performance of the Oracle and the Multi-feature beat tracker versus the number of committee members is presented. By comparing the improvements between the best ODF alone (*CSD*) when new members (given by the SFS method) are added to the committee, we find that after the sixth member is added, the performance is higher and statistically significant for the **AMLc** and **AMLt** measures. In Fig. 3(b). we show the improvement obtained by automatic selection between beat outputs using information gain and regularity. Table II presents the evaluation results of the best ODF mean performances of each of the genres of the *Dataset1360* per evaluation measure. There is no statistical difference between the results of the best three ODFs per genre. However some ODFs performed statistically worse than the others in these genres: Acoustic (*EF*), Afro-American (*PSF*), Classical (*BEF*, *EF*, *SFX*), Classical Solo (*SFX*), Electronic (*PSF*), Jazz (*EF*, *HF*, *MAF*), Rock/Pop (*PSF*), Samba (*HF*, *MAF*). Overall our results confirm the intuition that ODFs which are sensitive only to phase or harmonic changes are not the best choice for music genres with strong percussion, furthermore the *EF* is not a good choice for music without prominent percussion. Comparing the **AMLt** results of each onset detection function in *Dataset1360*, we find that 53% of the songs could be improved by using multiple ODF versus only using the single best performing onset detection function for this model, which led to an 11.6% average improvement.

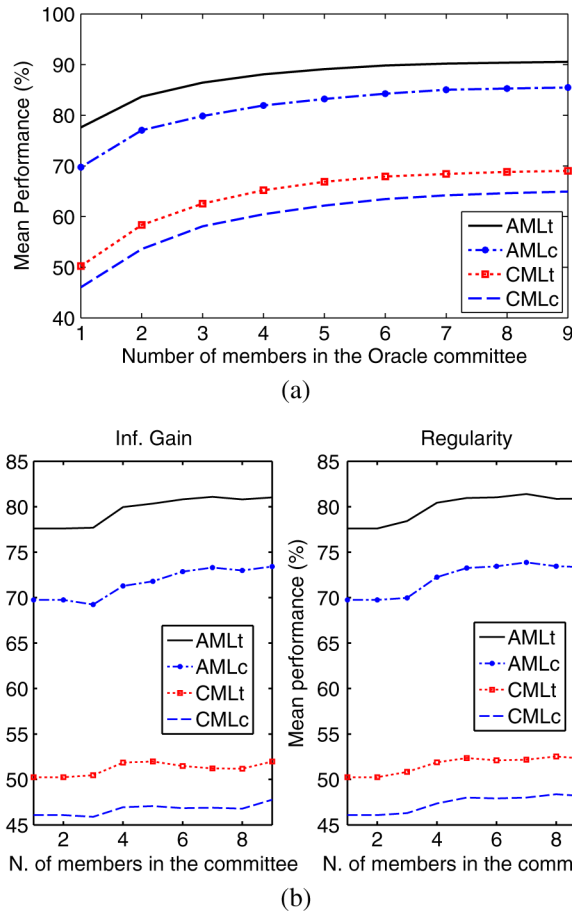

Fig. 3. (a) Oracle Mean Performance vs number of committee members. (b) Multi Feature (Inf Gain and Regularity) Mean Performance vs number of committee members.

TABLE II
MEAN PERFORMANCE (%) OF THE BEST FEATURE PER GENRE IN THE 1360 SONG DATASET

Genre	CMLc	CMLt	AMLc	AMLt
Acoustic	39.8 (*SFLF*)	45.6 (*SFLF*)	57.1 (*SFLF*)	67.6 (*SFLF*)
Afro-American	70.8 (*SFX*)	73.4 (*SFX*)	85.6 (*CSD*)	93.3 (*CSD*)
Balkan	19.6 (*EF*)	20.9 (*EF*)	77.4 (*SFLF*)	83.0 (*SFLF*)
Choral	8.8 (*PSF*)	13.9 (*PSF*)	16.4 (*HF*)	32.2 (*HF*)
Classical	38.7 (*BAS*)	47.0 (*BAS*)	53.9 (*BAS*)	67.5 (*HF*)
Classical Solo	31.0 (*BAS*)	33.3 (*BAS*)	66.1 (*BAS*)	73.6 (*BAS*)
Electronic	55.8 (*CSD*)	58.7 (*EF*)	81.6 (*SFX*)	83.6 (*SFX*)
Jazz	48.5 (*SFLF*)	54.9 (*CSD*)	68.1 (*SFLF*)	78.4 (*CSD*)
Rock/Pop	62.5 (*CSD*)	65.8 (*CSD*)	82.6 (*CSD*)	88.9 (*CSD*)
Samba	52.2 (*CSD*)	53.1 (*CSD*)	67.0 (*BEF*)	68.5 (*BEF*)

TABLE III
MEAN GROUND TRUTH PERFORMANCE (%) OF EACH BT ON *DATASET1360*. BOLD NUMBERS INDICATE BEST PERFORMANCES

Beat Tracker	CMLc	CMLt	AMLc	AMLt
Aubio (AUB) [40]	26.4	35.1	37.7	50.6
Beat.e [41]	36.1	42.2	61.6	74.0
Beatit (BIT) [42]	7.0	8.7	43.6	61.0
Beatroot (DIX) [43]	29.0	35.7	53.5	70.8
BeatUJAEN (BUJ) [44]	10.4	17.2	26.8	41.6
Boeck (BOE) [45]	31.5	43.5	42.2	58.7
BpmHistogram (BHI) [46]	13.8	21.6	34.4	57.3
Davies (DAV) [12]	46.8	50.8	69.3	75.9
Degara (DEG) [30]	46.0	50.2	69.9	77.7
Echonest[3]	31.7	36.3	52.0	59.8
Ellis (ELL) [29]	10.7	14.0	38.5	60.0
Gkiokas (GKI) [47]	41.5	47.1	62.7	72.7
Hainsworth (HAI) [48]	34.3	37.2	54.1	59.6
IBT off-line (IBT) [49]	32.5	36.9	64.0	73.8
Klapuri (KLA) [16]	**47.7**	**52.7**	69.8	77.7
Scheirer (SCH) [11]	21.2	34.5	30.4	49.0
Shine [50]	45.2	48.7	62.5	67.7
Stark (STA) [31]	41.7	47.3	61.6	71.0
MultiFt InfG	46.8	51.5	**72.9**	80.8
MultiFt Reg	**47.9**	**52.1**	**73.5**	**81.0**
MultiFt Essentia	46.2	50.7	71.9	80.4
5 BT Committee [14]	46.9	51.6	72.3	**81.4**
Oracle	64.9	69.0	85.5	90.5

[3] http://developer.echonest.com/

B. Comparison Results

In Table III, the mean accuracy of the different beat tracking algorithms is compared.

We present two configurations of the multi-feature beat tracker: the first with six committee members (*CSD*, *HF*, *EF*, *PSF*, *SFLF* and *BEF*) chosen by the SFS method with the information gain (Multi InfG) and regularity (Multi Reg) used in the selection step; and the second configuration (MultiFt Essentia) which is the C++ of the Multi-feature Information Gain (ZDG1) [39] submitted to the MIREX 2012 beat tracking task using *CSD*, *HF*, *EF*, *BEF* and *MAF*. This configuration is the released version of the Multi-feature beat tracker due to the disproportionately high computational cost of including the PSF ODF.

While the mean performance of all beat tracking systems is moderately low when using **CMLc** or **CMLt** (i.e., when the beats must be tapped at the annotated metrical level), performance naturally improves when we incorporate the additional, allowed metrical levels using **AMLc** and **AMLt**.

When comparing the proposed beat tracking algorithm with the reference systems, as shown in Table III, we see that the proposed method outperforms the reference methods in the mean value for all of the evaluation criteria. However, not all of the differences are statistically significant ($p < .05$). We find no significant differences between the proposed algorithm and the following reference methods with the DAV, DEG and KLA systems under **CMLc**, and then with the KLA system under **CMLt**.

When we compare MultiFt InfG and MultiFt Reg, which use a subset of six ODFs, with MultiFt Essentia which uses a subset of five ODFs, we do not find statistically significant differences. Furthermore we do not find any statistical difference compared to the committee system with five separate beat tracking algorithms (Beatroot, Degara, Ellis, Klapuri, IBT) as proposed in [14], which supports our hypothesis of diversity in input features being more important than using diverse beat tracking models.

C. Computation Time

Given the equivalent performance of our multi-feature approach with the 5 BT committee, we now address the differences in computational cost. To this end, we record the computation time for processing all the excerpts in *Dataset1360* (a total of 14 h 20 m 57 s). Simulations were conducted on a recent iMac (2.7 GHz Intel Core i5 with 8 GB RAM running MATLAB R2011b). For the 5 BT committee we found a total processing time of 1 h 56 m 06 s with the following breakdown per beat tracking component (DEG: 13 m 57 s, KLA: 45 m 42 s, ELL: 10 m 17 s, DIX: 28 m 43 s, IBT: 16 m 44 s and the MMA calculation took: 0 m 43 s). In comparison, the publicly available MultiFt Essentia implementation took just 19 m 17 s. To extract the beats in one minute of 44 kHz audio, the 5 BT committee takes approximately 8.1 s, where as Essentia algorithms takes only 1.3 s. In this sense, the use of a fast tracking algorithm (e.g. DEG as opposed to KLA) and an efficient C++ implementation can dramatically increase the speed with which the beats can be estimated.

D. Automatic Selection Results

To verify that using either Information Gain or Regularity as a selection mechanism provides a real improvement, we can compare the performance in Table III for our system with what happens if we make a random selection of the "best" beat tracking output per song. Running 100,000 trials we obtained mean performance of [40.6%, 45.3%, 64.6%, 73.3%] with variance [0.49, 0.50, 0.37, 0.32] for **CMLc**, **CMLt**, **AMLc**, **AMLt** respectively. The increase in performance we achieve using a structured, rather than random selection process is highly significant ($p < .00001$). However, the beat tracking accuracy from using either Information Gain or Regularity as a selection method falls well below the theoretical optimum of the Oracle system which can choose the best beat sequence per song, suggesting that automatic selection methods remains a profitable avenue for future work.

E. Mean Mutual Agreement and Confidence Threshold

To further explore the properties of the multi-feature beat tracker, we undertake an analysis of the *MMA* values. First, we seek to recreate the primary result from [14] which showed a high correlation between the *MMA* of the beat tracking committee and the mean performance of the committee against ground truth, the MGP. As shown in Fig. 4(a) we can see that the *MMA* (using Information Gain) is strongly correlated with the MGP of the committee using the set of ODFs. Thus we can confirm that disagreement between the beats of the committee is indicative of overall poor beat tracking accuracy and vice-versa. While we could not find a statistically significant difference in performance between the six member and nine member committees, we would like to explore the extent to which the mutual agreement changes based on the number of

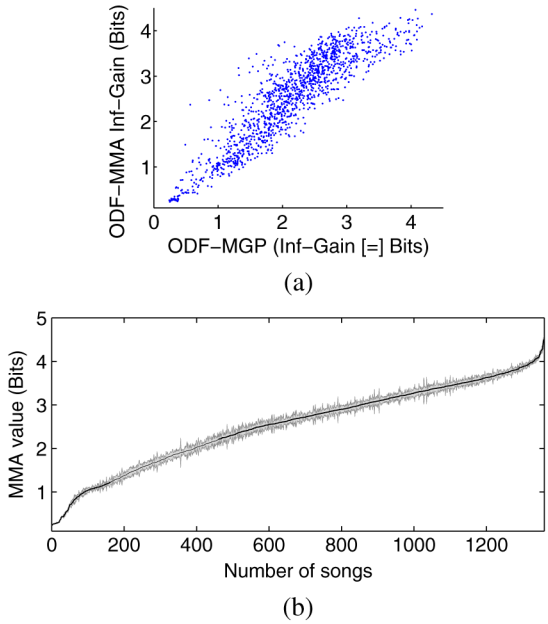

Fig. 4. (a) ODF mean mutual agreement (MMA) vs ODF Mean ground truth performance (MGP). (b) Error-bar of *MMA* calculated with 6 and more committee members vs songs, sorted by *MMA*(9 committee members).

committee members. To this end we show the range of observed *MMA* values obtained with committees of six, seven, eight and then nine members, in Fig. 4(b). As expected, we find very low variance in the *MMA* values obtained with committees of different sizes both when the mutual agreement is very low and likewise when it is very high. In this sense the variation in the size of the committee becomes apparent in the middle *MMA* range ($1.5 < \mathrm{MMA} < 3.5$ bits).

To complete our analysis of *Dataset1360*, we investigate whether we can use the *MMA* vs MGP correlation to automatically assign either high or low confidence to the estimated beats. Following [21], where a threshold of $MMA > 1.5$ bits was found to be indicative of acceptable beat tracking via a subjective listening test, we re-examine the beat tracking performance on songs with *MMA* above and below 1.5 bits. As shown in Table IV, we see that performance is far higher for excerpts where the *MMA* is above the threshold compared to below it. Of the 1360 excerpts in the dataset, we found 1126 (82.9%) were above it, for which the **AMLt** value $> 86\%$ for all configurations of the multi-feature beat tracker. While the beat tracking performance is lower for $MMA < 1.5$ bits this does not mean the multi-feature beat estimations cannot be accurate, merely that we do not have high confidence in the result. Likewise there will be cases with *MMA* above the threshold which are not accurate. These can arise when the beats are tapped at a meaningful metrical level, but one not included with in the set of allowed levels specified for **AMLt** [33].

F. MIREX Results

Thus far, all of our analysis has been *Dataset1360*, and while there is a wide diversity of musical genres and a large number of annotated files, we should acknowledge that the performance we observe might be slightly optimistic given access to the test data when choosing the committee members. Therefore, in addition to our own evaluation on the *Dataset1360*, we also report results

TABLE IV
MEAN SCORES (%) OF ORACLE, COMMITTEE OF 5 BEAT TRACKERS (5BT), MULTI-FEATURE BEAT TRACKER (MULTIFT) AND BEST MEAN PERFORMANCE BEAT TRACKER (BESTBT) FOR THE TWO SUBSETS OF *DATASET1360* DIVIDED BY AN *MMA* THRESHOLD OF 1.5 BITS

Beat Tracker	CMLc	CMLt	AMLc	AMLt	MMA
Oracle	70.4	72.8	91.8	94.4	>1.5
5BT	52.6	56.1	80.4	87.5	
MultiFt Reg	53.3	56.1	81.3	86.7	
MultiFt InfG	51.4	54.8	80.4	86.6	
MultiFt Essentia	51.7	55.0	80.0	86.3	
BestBt	53.5	57.4	77.6	84.0	
Oracle	38.5	50.8	54.8	71.8	<1.5
5BT	19.1	29.5	32.7	51.8	
MultiFt Reg	21.7	32.8	35.6	53.4	
MultiFt InfG	24.7	35.5	36.3	52.8	
MultiFt Essentia	19.2	29.8	32.7	51.5	
BestBt	19.9	29.8	31.7	47.0	

TABLE V
MIREX 2012 MEAN PERFORMANCE (%) AND THE BEST **AMLT** PERFORMANCE IN 2011,2010 AND 2009 IN MCK DATASET

Year	Beat Tracker	CMLc	CMLt	AMLc	AMLt
2012[4]	**ZDG2**	25.0	33.4	**51.8**	**66.7**
	GP3	23.7	33.7	49.3	66.5
	ZDG1	23.7	32.3	49.5	65.1
	GP2	23.3	32.3	48.6	64.9
	GKC2	25.8	32.9	51.0	64.2
	ODGR1	21.6	20.0	49.4	64.2
	FK1	22.3	35.1	41.5	63.3
	ODGR2	22.4	30.4	47.0	62.7
	KB1	17.5	29.9	35.9	60.2
	ODGR3	21.8	29.7	44.2	59.7
	FW4	23.7	34.5	42.4	59.1
	KFRO1	25.0	32.0	47.1	58.8
	ODGR4	20.0	28.3	41.4	58.2
	SB6	20.4	29.3	40.8	57.2
	FW3	22.5	34.1	39.2	57.0
	SB3	20.8	30.0	37.2	53.6
	GP4	19.6	30.4	35.2	52.5
	SB7	16.5	26.4	27.6	44.2
	SB4	14.2	24.0	24.4	42.1
	FW5	9.4	18.8	17.0	34.8
2011[5]	GP5	24.0	33.7	49.3	66.5
2010[6]	GP3	24.0	33.7	49.0	66.1
2009[7]	GP1	**26.0**	**35.5**	49.1	66.6

[4] http://nema.lis.illinois.edu/nema out/mirex2012/results/abt/mck/
[5] http://nema.lis.illinois.edu/nema out/mirex2011/results/abt/mck/
[6] http://nema.lis.illinois.edu/nema out/mirex2010/results/abt/mck/
[7] http:/music-ir.org/mirex/wiki/2009:Audio Beat Tracking Results

from the 2012 MIREX Audio Beat Tracking task, where we submitted two versions of our multi-feature beat tracker: ZDG1 and ZDG2 [39] which used *BEF*, *CSD*, *EF*, *HF* and *MAF* as committee members, and used the information gain and regularity selection methods respectively. The MIREX dataset is private and therefore can be considered as appropriate validation for our method.

In the Table V we show the 2012 MIREX results (sorted by **AMLt**) for the beat tracking task are presented, and also the best **AMLt** performers in 2011, 2010 and 2009 in the MCK dataset. The MCK dataset contains 160 30-sec. excerpts (WAV format) and has been used since the beginning of the MIREX

beat tracking evaluation in 2006. The excerpts in the MCK dataset each have an approximately stable tempo covering a wide distribution of tempi across all files, and a large variety of instrumentation and musical styles. About 20% of the files contain non-binary meters with a small number of examples contain changing meters.

As can be seen from the table, our multi-feature systems ZDG1 and ZDG2 perform competitively with the submitted algorithms for 2012 and those which have performed well in previous years. While the differences in performances are small between the most accurate systems we believe that the state of the art accuracy provided by our method shows that we have not over-fitted to *Dataset1360* to a degree which has adversely affected performance on the closed MIREX dataset.

V. Conclusions and Future Work

In this paper we presented a multi-feature beat tracking system. The main contribution of this research has been to demonstrate that the concept mutual agreement between beat tracking systems, presented in [14], can be extended to a committee of input features passed to a single beat tracking model, i.e. that there is sufficient diversity in input features to permit the use of a mutual agreement based selection method. In addition we make our beat tracking system freely available for re-use as part of the Essentia framework.

In a comprehensive comparison of the current state in beat tracking, we demonstrated that our system can outperform the current state of the art in beat tracking, and we have shown this both on the largest compiled beat tracking dataset, and in a closed MIREX evaluation. Beyond providing an accurate estimate of beat locations across a wide variety of musical styles, our system can also exploit the property that mutual agreement can indicate the level of confidence in the resulting beat locations. To this end, we showed that beat tracking performance is much more accurate when the mutual agreement between the committee members is high, compared to when the beat outputs lack consensus. In the wider context of re-use of our beat tracker, in particular for MIR applications which rely on beat-synchronous processing and assume the beats to be accurate, we can provide a guide over whether the beat locations ought to be trusted or not using *MMA*. Furthermore, the *MMA* can automatically indicate songs which are challenging for beat tracking, and therefore worthwhile for manual annotation as ground truth [32].

In terms of future work, we plan to investigate other input features which we could use as input to our multi-feature beat tracker, in particular those which can cater for more challenging beat tracking cases, e.g. for songs which lack clear percussive content. In addition we will also explore other techniques for automatically selecting between the committee of beat sequences (e.g. ROVER [51]), with the aim of closing the gap towards what is theoretically possible with an Oracle system and what can currently be achieved.

Acknowledgment

We thank the authors of the beat tracking algorithms and onset detection functions for making their code available. Finally, we would like to thank the anonymous reviewers for their valuable feedback and suggestions for improvement.

References

[1] F. Gouyon and S. Dixon, "A Review of automatic rhythm description systems," *Comput, Music J.*, vol. 29, no. 1, pp. 34–54, Mar. 2005.
[2] M. Goto and Y. Muraoka, "A beat tracking system for acoustic signals of music," in *Proc. 2nd ACM Int. Conf. Multimedia*, 1994, pp. 365–372.
[3] E. D. Scheirer, "Pulse tracking with a pitch tracker," in *Proc. Workshop Applicat. Signal Process. Audio Acoust.*, 1997.
[4] S. Dixon, "Beat induction and rhythm recognition," in *Proc. Australian Joint Conf. Artif. Intell.*, Perth, Australia, 1997, pp. 311–320.
[5] N. Degara, "Signal processing methods for analyzing the temporal structure of music exploiting rhythmic knowledge," Ph.D. dissertation, Univ. of Vigo, Vigo, Spain, 2011.
[6] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proc. 10th Int. Soc. for Music Inf. Retrieval Conf.*, 2009, pp. 231–236.
[7] M. Levy and M. B. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 2, pp. 318–326, Feb. 2008.
[8] S. Ravuri and D. P. Ellis, "Cover song detection: From high scores to general classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 65–68.
[9] J. A. Hockman, J. P. Bello, M. E. P. Davies, and M. D. Plumbley, "Automated Rhythmic Transformation of Musical Audio," in *Proc. 11th Int. Conf. Digital Audio Effects*, Espoo, Finland, 2008, pp. 177–180.
[10] A. Robertson and M. Plumbley, "B-Keeper: A beat-tracker for live performance," in *Proc. Int. Conf. New Interfaces for Musical Expression (NIME)*, New York, NY, USA, 2007, pp. 234–237.
[11] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.
[12] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 3, pp. 1009–1020, Mar. 2007.
[13] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *J. New Music Res.*, vol. 36, no. 1, pp. 1–16, Mar. 2007.
[14] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2539–2548, Nov. 2012.
[15] N. Collins, "Towards a style-specific basis for computational beat tracking," in *Proc. 9th Int. Conf. Music Percept. Cognition*, 2006, pp. 461–467.
[16] A. P. Klapuri, A. J. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.
[17] M. E. P. Davies and M. D. Plumbley, "Comparing mid-level representations for audio based beat tracking," in *Proc. DMRN Summer Conf.*, Glasgow, U.K., 2005.
[18] A. M. Stark, "Musicians and machines: Bridging the semantic gap in live performance," Ph.D. dissertation, Queen Mary, Univ. of London, London, U.K., 2011.
[19] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A Tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.
[20] F. Gouyon, S. Dixon, and G. Widmer, "Evaluating low-level features for beat classification and tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 1309–1312.
[21] 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 *Proc. 13th Int. Soc. Music Inf. Retrieval Conf.*, Porto, Portugal, 2012, pp. 157–162.
[22] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in *Proc. Int. Conf. Music Inf. Retrieval. ISMIR*, Philadelphia, PA, USA, 2008, pp. 653–658.
[23] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *J. Audio Eng. Soc.*, vol. 51, no. 4, pp. 226–233, 2003.
[24] P. Masri, "Computer modeling of sound for transformation and synthesis of musical signal," Ph.D. dissertation, Univ. of Bristol, Bristol, U.K., 1996.
[25] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proc.13th Int. Soc. Music Inf. Retrieval Conf. ISMIR*, Porto, Portugal, 2012, pp. 49–54.

[26] C. Duxbury, J. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. 6th Conf. Digital Audio Effects (DAFx)*, London, U.K., 2003.
[27] M. E. P. Davies, M. M. D. Plumbley, and D. Eck, "Towards a musical beat emphasis function," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2009, pp. 61–64.
[28] S. Hainsworth and M. Macleod, "Onset detection in musical audio signals," in *Proc. Int. Comput. Music Conf.*, Singapore, 2003, pp. 136–166.
[29] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, Mar. 2007.
[30] N. Degara, E. A. Rua, A. Pena, S. Torres-Guijarro, M. E. P. Davies, and M. D. Plumbley, "Reliability-informed beat tracking of musical signals," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 1, pp. 290–301, Jan. 2012.
[31] A. M. Stark, M. E. P. Davies, and M. D. Plumbley, "Real-time beat-synchronous analysis of musical audio," in *Proc. 12th Int. Conf. Digital Audio Effects (DAFx-09)*, 2009, pp. 299–304.
[32] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, "On the automatic identification of difficult examples for beat tracking: Towards building new evaluation datasets," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 89–92.
[33] M. E. P. Davies, N. Degara, and M. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Centre for Digital Music, Queen Mary Univ. of London, C4DM-TR-09-06, 2009, Tech. Rep..
[34] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Measuring the performance of beat tracking algorithms using a beat error histogram," *IEEE Signal Process. Lett.*, vol. 18, no. 3, pp. 157–160, 2011.
[35] M. Marchini and H. Purwins, "Unsupervised analysis and generation of audio percussion sequences," in *Exploring Music Contents*. Berlin, Heidelberg, Germany: Springer, 2011, pp. 205–218.
[36] F. Gouyon, "A Computational approach to rhythm description," Ph.D. dissertation, Pompeu Fabra Univ., Barcelona, Spain, 2005.
[37] S. Hainsworth, "Techniques for the automated analysis of musical audio," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2004.
[38] F. Gouyon and P. Herrera, "Determination of the Meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Proc. 114th Conv. Audio Eng. Soc.*, 2003.
[39] J. R. Zapata, M. E. P. Davies, and E. Gomez, "MIREX 2012: Multi Feature beat tracker (ZDG1 AND ZDG2)," in *Music Inf. Retrieval Eval. eXchange (MIREX)*, Porto, Portugal, 2012.
[40] P. M. Brossier, "Automatic annotation of musical audio for interactive systems," Ph.D. dissertation, Queen Mary Univ. of London, London, U.K., 2006.
[41] F. Krebs and G. Widmer, "Audio beat tracking evaluation: Beat.e.," in *Music Inf. Retrieval Evaluation eXchange (MIREX)*, Porto, 2012.
[42] J. Bonada and F. Gouyon, "Beatit, mtg.upf.edu, internal software," 2006.
[43] S. Dixon, "Evaluation of the audio beat tracking system BeatRoot," *J. New Music Res.*, vol. 36, no. 1, pp. 39–50, 2007.
[44] R. Mata-Campos, F. J. Rodriguez-Serrano, P. Vera-Candeas, J. J. Carabias-Orti, and F. J. Canadas-Quesada, "Beat tracking improved by am sinusoidal modeled onsets - Mirex 2010," in *Music Inf. Retrieval Evaluation eXchange (Mirex)*, 2010.
[45] S. Böck and M. Schedl, "Enhanced beat tracking with context-aware neural networks," in *Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11)*, 2011, pp. 135–139.
[46] E. Aylon and N. Wack, "Beat detection using PLP," in *Music Inf. Retrieval Evaluation eXchange (MIREX)*, 2010.
[47] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stajylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 421–424.
[48] S. W. Hainsworth and M. Macleod, "Particle filtering applied to musical tempo tracking," *J. Adv. Signal Process.*, vol. 15, pp. 2385–2395, 2004.
[49] J. Oliveira, M. E. P. Davies, F. Gouyon, and L. P. Reis, "Beat tracking for multiple applications: A multi-agent system architecture with state recovery," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 10, pp. 2696–2706, Oct. 2012.
[50] M. Khadkevich, T. Fillon, G. Richard, and M. Omologo, "A probabilistic approach to simultaneous extraction of beats and downbeats," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 445–448.
[51] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Autom. Speech Recogn. Understand.*, 1997, pp. 347–354.

José R. Zapata is an assistant professor at the Department of Information and Communication Technologies, Universidad Pontificia Bolivariana (UPB). He received the B.Eng. degree in electronic in 2003 and the M.Sc. degree in Telecommunications in 2008, both from UPB, Medellin, Colombia. In 2013, he completed his Ph.D. at the Universitat Pompeu Fabra (UPF), on the topic of Automatic rhythm description systems. His research interests include beat tracking and rhythm analysis, evaluation methods and MIR applications.

Matthew E. P. Davies received the B.Eng. degree in computer systems with electronics from Kings College London, U.K., in 2001 and the Ph.D. degree in electronic engineering from Queen Mary University of London (QMUl), U.K., in 2007. From 2007 until 2011, he was a Postdoctoral Researcher in the Centre for Digital Music, QMUl. In 2011 he joined the Sound and Music Computing Group at INESC TEC in Porto, Portugal. His research interests include beat tracking and rhythm analysis, evaluation methods, music therapy, and creative-MIR applications.

Emilia Gómez is a Postdoctoral Researcher and assistant professor at the Music Technology Group (MTG), Department of Information and Communication Technologies, Universitat Pompeu Fabra (UPF). She received the B.Eng. degree in Telecommunication specialized in Signal Processing at Universidad de Sevilla. Then, she received a DEA in Acoustics, Signal Processing and Computer Science applied to Music at IRCAM, Paris. In 2006, she completed her Ph.D. in Computer Science and Digital Communication at the UPF, on the topic of Tonal Description of Music Audio Signals. Her research interests include melodic and tonal description of music audio signals, computer-assisted music analysis and computational ethnomusicology.