2011-01-01

Tempo Detection Using a Hybrid Multi-band Approach

Mikel Gainza
Technological University Dublin, Mikel.Gainza@tudublin.ie

Eugene Coyle
Technological University Dublin, Eugene.Coyle@tudublin.ie

Recommended Citation

Gainz, M. & Coyle, E. (2011) Tempo Detection using a hybrid multi-band approach. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, vol: 19, issue: 1, pages: 57 - 68. doi:10.1109/TASL.2010.2045182

Tempo Detection Using a Hybrid Multiband Approach

Mikel Gainza and Eugene Coyle

***Abstract*—In this paper, a novel tempo detection system is presented, which suggests the use of a hybrid multiband decomposition. The model tracks the periodicities of different signal property changes that manifest within different frequency bands by using the most appropriate onset/transient detectors for each frequency band. In addition, the proposed system applies a novel method to weight tempo candidates. Each contribution is evaluated by comparing the presented system against existing approaches using three different databases that comprises 1638 songs. These databases include the two publicly available database of songs used in the tempo evaluation contest of ISMIR 2004. These songs are used in order to compare the proposed approach against four recent existing approaches and also against the participants of the tempo detection contest of ISMIR 2004. The results show that the presented approach provides an improvement over existing techniques.**

***Index Terms*—Onset detection, periodicity detection, rhythm description, tempo detection.**

I. Introduction

RHYTHM is characterized by patterns of musical units that occur at different hierarchical metrical levels. The rhythmic units that occur at the primary metrical level are called beats and the rate of repetition of these beats provides the tempo of a piece of music, which is expressed in beats per minute (bpm). In staff notation, the primary metrical level is given by the denominator of the annotated time signature and the annotated tempo provides the duration of the beats in that metrical level. As an example, a song annotated with a time signature 4/4 and a tempo equal to 120 bpm will have its primary metrical level at the crotchet level and the constituent beats will have a duration of $60/120 = 0.5$ s. A less formal interpretation is understanding tempo as the rate at which humans tap along their feet while listening to music. However, the annotated tempo does not always correspond to the perceived beats by humans [1], who can perceive the tempo of the same song differently. Nevertheless, as [2] indicates, these tempo deviations generally correspond to a rhythmic perception at different metrical levels. Thus, deviation of the tempo in factors 2 or 1/2 generally occur for duple meter music and deviations in factors 3 or 1/3 occur for both compound meters and triple meters. The perceptual distribution of tempo of different groups of listeners is investigated in [3], where differences from a "predicted" tempo range are explained by the existence of perceptual periodic dynamic accents in the musical excerpt [3], [4].

The vast amount of existing research in this area is explained by the large variety of applications derived from the automatic detection of the tempo. As an example, music information retrieval systems allow retrieving songs which have similar tempo to a specific query. Other applications that use tempo information include automatic playlist generation, music similarity computations, beat tracking algorithms, music performance and style research, DJ mixing applications, and audio track synchronization.

This paper is organized as follows. First, Section II describes existing research in the area of tempo detection. In addition, areas of potential improvement of existing approaches are identified. Following this, Section III introduces the proposed tempo detection approach. Next, a set of results obtained by evaluating the presented approach using three different databases of musical signals are presented in Section IV, which is followed by a discussion of the obtained results in Section V. Finally, conclusions and directions for future work are given in Section VI.

II. Existing Tempo Detection Research

Existing tempo detection methods generally share a similar framework [2]. First, the audio is converted into a downsampled representation where the frames around onset times are emphasized by generating an *Onset Detection Function (ODF)*,[1] which tracks different signal property changes. Next, the existing periodicities of the ODF are extracted, which results in the generation of a *Periodicity Detection Function (PeDF)*. Finally, the $PeDF$ is *postprocessed* in order to extract the periodicity that corresponds to the perceived tempo.

The choice of ODF, $PeDF$ and postprocessing techniques vary significantly between existing tempo detectors. As an example, the ODF used in [5] tracks sharp energy changes in the signal, [6] attempts to model the human auditory system, the system used in [7] tracks complex spectral changes, and the spectral flux is used in [8]. The autocorrelation function is the most widely utilized $PeDF$ [8]–[10]. Other periodicity detection functions include comb filters [11], [12], methods based on spectral analysis [6], [8] or phase-preserving autocorrelation function [13]. The postprocessing technique utilized to

Manuscript received April 20, 2009; revised December 23, 2009; accepted January 14, 2010. Date of publication March 11, 2010; date of current version October 01, 2010. This work was supported in part by Enterprise Ireland under Project IMAAS, CFTD/06/220. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dan Ellis.

The authors are with the Dublin Institute of Technology, Audio Research Group, Dublin 2, Ireland (e-mail: mikel.gainza@dit.ie; eugene.coyle@dit.ie).

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/TASL.2010.2045182

[1]The term Onset Detection Function (ODF) refers to a function whose peaks ideally coincide with onset times. In the context of a tempo detector, it does not necessary imply musical onset times being extracted.

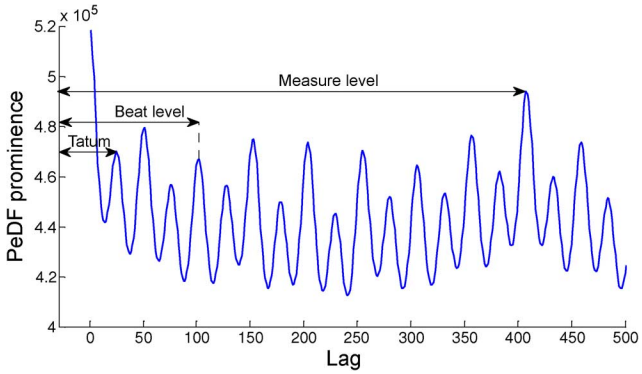

Fig. 1. $PeDF$ of an excerpt of song "Do your best" by "Femi Kuti."

extract the tempo from the $PeDF$ also varies between existing approaches, where simple methods such as getting a maximum in the $PeDF$ [14], investigating hierarchical meter relations between peaks [13], dynamic programming techniques that evaluate tempo hypotheses [15] or more complex probabilistic models have been utilized [6], [11]. An example of the different metrical levels that can be estimated and utilized within a $PeDF$ is illustrated Fig. 1. In this figure, the autocorrelation function of an ODF of an excerpt of song "Do your best" by "Femi Kuti" is shown, in which three different metrical levels are manually labeled. The periodicity corresponding to the beat period is located at $\text{lag} = 101$, which in this example corresponds to a tempo equal to 102 bpm.[2] As can be seen in Fig. 1, extracting the most prominent periodicity in the $PeDF$ within a certain tempo range will not necessarily lead to correct tempo estimation.

Alternative approaches that do not adhere to the general tempo detection framework include methods that estimate onset times before periodicity detection [5], [15], [16]. Other methods exploit the structured and repetitive nature of certain music types by building a similarity matrix of the music signal [17], [18]. The resulting matrix diagonals are processed and the diagonal which corresponds to the song's inter-beat interval (IBI) will contain a higher degree of music similarity.

Multiband approaches have been widely used in tempo detection systems. In [12], Scheirer splits the signal into six frequency bands. Following this, the periodicity of the amplitude envelopes of the filterbank outputs are extracted by using a bank of comb filter resonators. The output of the resonators is summed across the frequency bands and the frequency of the resonator with most energy will correspond to the tempo of the piece of music. Klapuri *et al.* base their meter detection system on Scheirer's model by using four "accent bands," which combine the loudness differences of 36 frequency bands [11]. The model uses a comb filter bank to seek periodicities in three different metrical levels (tatum, beat and bar) in each of the four accent bands. Then, a probabilistic model of the dependencies and temporal relations between the three metrical levels is performed. The model does not explicitly calculate the tempo. However, the system is modified in [19] in order to calculate the tempo as the median of the estimated beat positions. In [10], the three most prominent peaks of the eight normalized band autocorrelation functions are used to combine periodicities across bands. The multiband method presented by Uhle [9] tracks periodicities in two different time ranges. First, the tatum is calculated within a short time range. Then, periodicity multiples of the tatum that fit predefined rhythmic templates will be used in order to calculate the tempo in a larger time range. Another novelty of this method is the prior segmentation of the signal into regions of audio similarity under the assumption that a new region (e.g., a new verse or chorus) might trigger a tempo change. Ellis calculates the tempo by obtaining the autocorrelation function of an onset detection signal, which is calculated by using a log-magnitude 40 channel mel-frequency spectrogram [14]. The periodicity detection is performed after summing across frequency bands. As [19] states, the difference between calculating the $PeDF$ before or after summing across bands lies in the fact that the former will only detect periodicities present in the analyzed band and the latter will emphasize periodicities present in all bands. Alonso also used a multiband approach, where spectral methods and the autocorrelation function were used in order to obtain tempo hypotheses. Following this, dynamic programming techniques were used in order to find the tempo hypothesis that best explains a list of observed onsets [15].

The literature presented above gives an overview of the main tempo detection methods. For a more extended review, readers can refer to [2] and [20]. In addition, methods that participated in ISMIR 2004 and MIREX 2006 tempo detection contests are described in [19] and [21], respectively. The approach used by Klapuri in [11] participated in both contests winning on both occasions. The other MIREX tempo evaluation contest was organized in 2005 and won by the approach presented by Alonso in [15].[3] It should be noted that both Alonso's and Klapuri's methods use a multiband decomposition.

As previously discussed, the general method of tempo detection lies in the identification of the ODF periodicity that corresponds to the music tempo. Consequently, the generation of an accurate ODF is of crucial importance. The choice of the onset detector significantly varies between existing tempo detection models. In [22], Davies compares the performance of seven different onset detectors, including Klapuri's and Scheirer's onset detectors [11], [12], for the purpose of tempo detection and beat tracking. The results show that the spectral complex change onset detection method, presented in [23], is the most suitable representation for tempo detection. In [24], Gouyon *et al.* compare the use of 172 different low-level acoustical features as a front end to a beat tracking system The results show that the spectral complex change feature provides the best performance overall for that task. This onset detector is used by Davies *et al.* in the tempo detection model presented in [7], which justifies which justifies the choice of Davies *et al.* approach as the model to base our approach on.[4] However, in contrast to Alonso's and Klapuri's methods [11], [15], Davies

[2]Calculated using (16), where $H = 256$ samples and $fs = 44100$ Hz.

[3]Alonso and Klapuri did not participate in ISMIR 2004 and MIREX 2005 tempo detection contests, respectively. Both Alonso and Klapuri participated in MIREX 2006 tempo detection contest [21].

[4]Davies *et al.*'s method participated in MIREX 2006 tempo detection contest, finishing in second position. Klapuri and Alonso were first and third, respectively [21].

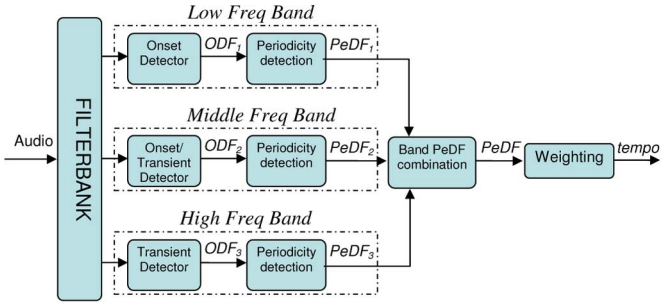

Fig. 2. Proposed tempo detection system.

et al.'s method does not use a multiband approach. In this paper, the impact of adapting Davies *et al.*'s tempo detection method into a multiband decomposition, which also uses the spectral complex change onset detector is investigated. In addition, a novel method of using different onset detection algorithms within each frequency band is presented in the following section. The proposed method attempts to exploit the advantages of tracking different signal properties at different frequency ranges. Furthermore, a different strategy to weight the resulting cross-band $PeDF$s is also introduced in the proposed system.

III. Proposed System

Fig. 2 illustrates the different blocks that form the tempo detection system proposed here. First, a multi-band decomposition is utilized, which splits the incoming audio signal into three different frequency bands. Following this, the model attempts to use the most appropriate onset/transient detection method in each band. This is performed by exploiting the different acoustic properties of each frequency band with a different onset detector. Next, the existing band periodicities are extracted by building a $PeDF$ in each band. Following this, the band $PeDF$s are combined into a single representation. Next, the combined $PeDF$ is postprocessed by using a weighting function. Finally, the tempo is extracted from the weighted $PeDF$.

Section III-A introduces the multiband decomposition used in the presented approach. A brief description of the onset/transient detectors is given in Section III-B, which includes a discussion of the suitability of the onset/transient detectors in each frequency band. Following this, the characteristics of the hybrid multiband configuration are given in Section III-C. Then, the periodicity detection method is described in Section III-D. Finally, a description of the suggested weighting method is given in Section III-E.

A. Multiband Decomposition

The presented multiband tempo detection system splits the audio signal into three different frequency bands. The choice of the band cutoff frequencies is motivated by the different activity of certain instruments at different frequency regions. The different frequency ranges are given as follows.

- *Low-frequency band (LFB)*: frequency range: [0–200 Hz]. Existing periodicities resulting from the presence of a bass line or percussive instruments such as a snare or a kick drum will be present in this low-frequency band.
- *Middle-frequency band (MFB)*: frequency range: [200–5000 Hz]. This band range overlaps with a large number of instrument frequency ranges. Thus, this band will contain a large amount of energy and active frequency components. The chosen band range roughly covers the fundamental frequencies of a wide range of instruments.
- *High-frequency band (HFB)*: frequency range: $[5000 - fs/2$ Hz], where fs corresponds to the sampling rate. The presence of percussive instruments in the recording results in transient signals spreading over the entire frequency range. Due to the low presence of nonpercussive instruments in this band, transients will be more llocalizedin this band.

B. Onset/Transient Detection Function

As described in the previous section, a large number of different onset detection functions have been used within tempo detection systems. In the presented tempo detection system, the combination of the spectral complex change onset detection method, [23], and a transient detection method presented in [25] is suggested. In both methods, the frequency evolution over time is obtained using the short-time Fourier transform (STFT), which is calculated using a Hanning window and an FFT length N. The STFT is given by

$$X(n,k) = \sum_{m=0}^{L-1} x(m+nH)w(m)e^{-j(2\pi/N)km} \quad (1)$$

where $w(m)$ is the window that selects an L length block from the input signal $x(m)$, n is the frame number and H is the hop length in samples.

A brief description of the chosen onset/transient methods and its suitability to track periodicities in the above frequency bands is given as follows.

- *Spectral Complex change onset detection method (SC)* [23]: As described in Section II, this method was identified by [22] and [24] as a very suitable representation for tempo extraction. The method emphasizes onsets in the ODF by tracking energy changes in the magnitude spectrum and unexpected deviations in the phase spectrum (e.g., a pitch change). Thus, measurements and predictions of both energy and phase of frequency bins are calculated in order to generate a measured complex number S_k and a predicted complex number $\hat{S}k$, respectively, of each frame frequency bin. The difference between the predicted and measured complex number for bin k of a given frame n is calculated as follows:

$$\Gamma_k(n) = \left\{ \left[\Re(\hat{S}_k) - \Re(S_k) \right]^2 + \left[\Im(\hat{S}_k) - \Im(S_k) \right]^2 \right\}^{1/2} \quad (2)$$

where $\Re$ and $\Im$ are the real and imaginary parts, respectively.

The onset detection function frame is then generated by summing across frequency bin spectral complex changes as follows [23]:

$$ODF(n) = \sum_{k=1}^{N/2} \Gamma_k(n). \quad (3)$$

The use of the SC method in the three frequency bands can be seen as the result of turning the method presented in [7] into a multiband tempo detection method. The impact of using such configuration is evaluated in Section IV. The SC will effectively track energy changes in the LFB. In addition, the phase part of the complex number prediction facilitates the detection of slow onsets, such as a flute onset, and common onset energy changes occurring in the MFB. However, low-energy transients will be more difficult to track by using the SC in the HFB.

- *Transient detection method (TD)* [25]: This method, which has not yet been utilized within a tempo detection model, tracks the occurrence of broadband signals. This is performed by solely counting the number of bins that show an energy increase between consecutive frames larger than a threshold in dB [25]. The transient change for bin k of a given frame n is calculated as follows:

$$T(k,n) = 20\log_{10}\frac{X(k,n)}{X(k,n-1)}. \tag{4}$$

Then, the onset detection frame is calculated by counting the number of bins that reach the threshold $Thresh$ as follows:

$$ODF(n) = \sum_{\mathrm{k}=1}^{N/2}\begin{bmatrix}1, & \text{if} \quad T(k,n) > Thresh\\ 0, & \text{elsewhere}\end{bmatrix}. \tag{5}$$

Due to the low number of bins that comprise the LFB, the TD will not be a suitable method for this band. The TD will track percussive occurrences in the MFB. Since the energy content of the signal does not play an important role in the TD method, it will also be effective in tracking transients in the HFB. Thus, even if the energies of the constituent bins of a transient signal are low, the method will effectively track a new occurrence if the transient spreads over the HFB range.

C. Hybrid Multiband Configuration

As can be derived from the description of the three frequency bands, different signal property changes manifest at different frequency bands. Consequently, the use of the most appropriate onset/transient detection method in each frequency band depending on the acoustic properties of each band should improve the performance of a tempo detection model. The advantages of both transient and complex detectors are combined together into a hybrid model.

The configuration of the suggested hybrid multiband configurations $Hyb1$ and $Hyb2$ is shown in Table I. In the LFB, onset energies can span over several consecutive frames. In this case, the SC is a more suitable method to track energy changes than the TD and will be used in both hybrid configurations. In contrast, the use of TD in the HFB will ensure that existing broadband low energy transients will be accurately tracked. The method suitability in the MFB will change depending on the music type; singing solos or recordings with presence of slow onset instruments will benefit from the use of the SC (see $Hyb1$ method in Table I). In contrast, the TD will be more appropriate to detect percussive transients within complex polyphonies (see $Hyb2$ method in Table I).

TABLE I
PROPOSED HYBRID MULTIBAND CONFIGURATIONS

	Bands		
Configuration name	***LFB***	***MFB***	***HFB***
Hyb1	*SC*	*SC*	*TD*
Hyb2	*SC*	*TD*	*TD*

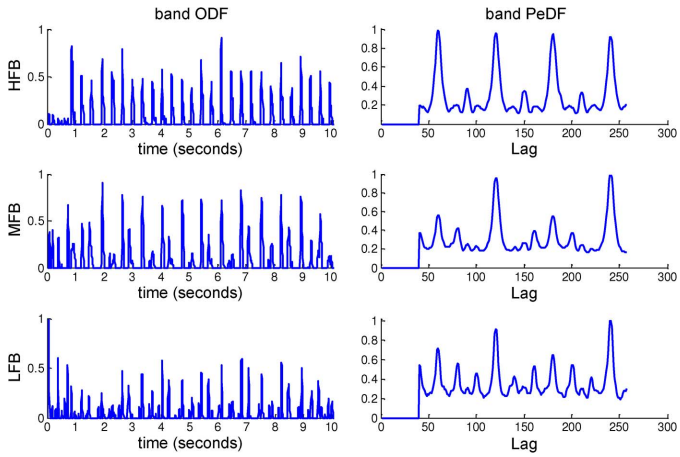

Fig. 3. Band ODFs (left column) and $PeDF$s (right column) of the three frequency bands LFB (bottom row), MFB (middle row), and HFB (top row) using the $Hyb1$ method in an excerpt of song "Big Time Operator" by "Big Band Batty Bernie."

The calculation of both onset/transient detection functions is performed by using a STFT with a frame length equal to 512 samples and 50% overlapping between consecutive frames. In the TD, the threshold $Thresh$ is set to 6 dB. Then, the band ODFs are postprocessed in order to generate a more smooth detection signal. First, the ODFs are processed by applying a third-order Butterworth IIR filter with cutoff frequency equal to $0.2 * (fs/2)$ to the signal. This filter is processed in forward and backward directions. Thus, it will affect peaks and decays in a similar manner. Next, as in [7], a moving mean threshold is calculated using windows with a duration equal to 0.2 s long. Then, the threshold is subtracted from the IIR filtered ODF, which has the effect of removing less significant peaks [7].

As an example, the left column of Fig. 3 depicts the band ODFs generated using $Hyb1$ method in a 10-s excerpt of Jive song "Big Time Operator" by "Big Band Batty Bernie." It can be seen that percussive transients are well localized using the TD in the HFB.

D. Periodicity Detection

As can be seen in Fig. 2, existing band periodicities are tracked by generating a $PeDF$ in each band. This is performed by using the widely utilized autocorrelation function r within each band ODF. Existing periodicities in the lag range $D = \{minLag \ldots maxLag\}$ are tracked, where $minLag$ and $maxLag$ correspond to the beat period (in frames) of a tempo equal to 250 bpm and 40 bpm, respectively.

$$r(D) = \sum_{n=1}^{Lo-D} ODF(n)ODF(n+D) \tag{6}$$

where Lo and n correspond to the length of the onset detection function ODF and the frame number, respectively.

In [7], the r periodicities are tracked by using comb filter templates, which correspond to a sum of weighted delta functions located at different periodicities. For each periodicity D, a comb filter template extracts values in regions R of the autocorrelation function centred at $D*l$, where $l = \{1 \dots 4\}$. The width of the regions is scaled proportionally to l. Thus, each region R contributes equally to the comb filter template

$$PeDF_i(D) = \sum_{l=1}^{4} \sum_{v=1-l}^{l-1} \frac{r_i(D^*l+v)}{2l-1}. \tag{7}$$

In the proposed multiband approach, a method based on (7) is adopted.[5] In order to better track deviations from perfect periodicity multiples, the maximum value of each region R within r is used instead of the delta functions. The ith band $PeDF$ is calculated as follows, where more weight to low multiples of D is given

$$PeDF_i(D) = \sum_{l=1}^{4} \frac{\max\left(r_i(D^*l+R)\right)}{l} \dots \text{where}$$
$$R = \{1-l \dots l-1\} \tag{8}$$

where $max(x)$ corresponds to the maximum value within region x.

The right column of Fig. 3 depicts the $PeDF$ of the three band ODFs shown on the left column of Fig. 3. It can be seen that the most prominent periodicity in each band varies. In the HFB, periodicities of existing percussive transients are accurately tracked, where the influence of other instrument periodicities is reduced.

In Fig. 2, it can be seen that the band $PeDF$s are combined into a single $PeDF$. This is achieved by summing the three maximized band $PeDFs$ as follows:

$$PeDF(D) = \sum_{i=1}^{3} \frac{PeDF_i(D)}{\max(PeDF_i)}. \tag{9}$$

E. Weighting Functions

Finally, as can be seen in Fig. 2, the combined $PeDF$ is weighted in an effort to reduce the number of double and half tempo estimations. The general method weights the $PeDF$ by a function W that gives different weight to each beat periodicity candidate

$$PeDF'(D) = PeDF(D) \times W(D). \tag{10}$$

Existing approaches generate the function W by using statistics derived from commonly used tempo annotations in popular music. As an example, Klapuri *et al.* use a lognormal distribution of the manually annotated database of tempos used in [11]. The function used in [7] weights the $PeDF$ by using a Rayleigh function as follows:

$$W(D) = \frac{D}{D_0^2} e^{\frac{-D^2}{2D_0^2}} \tag{11}$$

[5]The advantage of the use of both comb filter templates ((7) and (8)) is evaluated in Section IV.

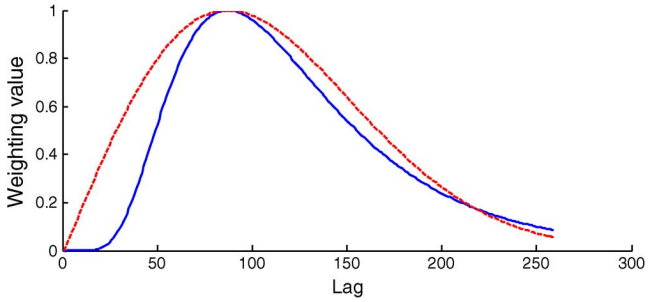

Fig. 4. Gaussian and Rayleigh weighting functions depicted in solid and dotted line, respectively.

where D_0 is the center of the function.

In [14], Ellis uses a Gaussian weighting function, which is given by

$$W(D) = \exp\left\{-\frac{1}{2}\left(\frac{\log_2 D/D_0}{\theta}\right)\right\} \tag{12}$$

where θ corresponds to the width of the function in octaves [14].

Both weighting functions are shown in Fig. 4. As in [7] and [14], the functions use a $D_0 = 0.5$ s, which corresponds to a tempo equal to 120 bpm. As in [14], the variable θ is set at 1.4 octaves.

The left plot in Fig. 4 shows the combined $PeDF$ of a song example. From the figure, the periodicity of the beat and its half subdivision are denoted as B1 and B2, respectively. In addition, the periodicity of the triplets played by the drummer are denoted as T1 and T2. In this case, the strongest periodicity in $PeDF$ is located at B1 and is denoted as M. The weighted $PeDF$ is shown on the right plot of Fig. 5, which illustrates that by weighting the $PeDF$ there is a potential risk of substantially weighting nonmultiples of the annotated tempo. It can be seen that the maximum in the weighted $PeDF$ is now located at a 2/3 subdivision of B1, which corresponds to the periodicity of the triplet subdivision T2. In order to overcome this problem, the following weighting technique is proposed.

- First, the most prominent periodicity M in the combined $PeDF$ is estimated within a wider lag range $D2 = \{minLag \dots maxLag * 2\}$. Extending the range of periodicities allows the estimation of prominent periodicities located at higher metrical levels such as the bar length.
- Following this, the $PeDF$ is weighted using a Rayleigh function [see (10) and (11)]. Next, only values in the weighted $PeDF'$ within specific ranges $R1_i$ and $R2_j$ centred at multiples and integer fractions, respectively, of M are used. It should be noted that even though M is estimated within a wide lag range $D2$, the tempo of the piece of music will be limited to the lag range $D = \{minLag \dots maxLag\}$ as stated in Section IV-D.
- In order to allow deviations from perfect periodicities, a deviation from each region center is allowed. Thus, the ith range $R1$ is given by

$$R1_i = \{M \times i - 2 \dots M \times i + 2\}$$
$$\text{where, if } M < maxLag \rightarrow i = \left\{1 : \left\lfloor \frac{maxLag}{M} \right\rfloor\right\}$$
$$\text{else} \rightarrow i = 1 \tag{13}$$

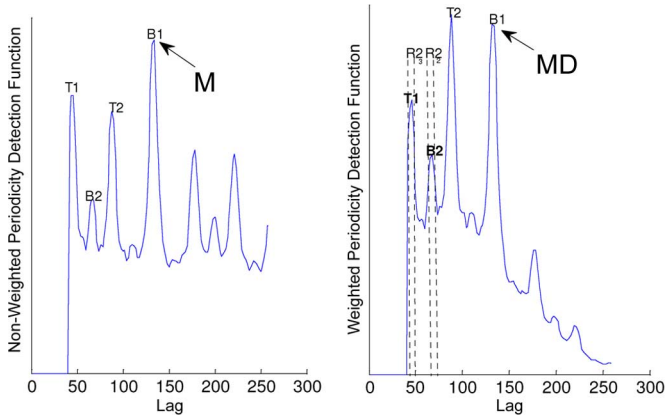

Fig. 5. Example of the use of the proposed weighting method. T1 and T2 correspond to the periodicities of triplets played by a drummer. B1 and B2 are the beat period and half beat period, respectively.

where $\lfloor\ \rfloor$ denotes nearest integer towards zero.
The jth range $R2$ is given by

$$R2_j = \left\{ \frac{M}{j} - 2 \ldots \frac{M}{j} + 2 \right\}$$
$$\text{where } j = \left\{ 1 : \left\lfloor \frac{M}{\text{minLag}} \right\rfloor \right\}. \quad (14)$$

- Finally, the periodicity MD that corresponds to the most prominent value within the above regions will be used in order to calculate the global tempo

$$\text{MD} = \arg_D \max \left(PeDF'(R) \right)$$
$$\text{where } R = \{R1_i, R2_j\} \quad (15)$$
$$\text{tempo} = \frac{fs \times 60}{MD \times H} \quad (16)$$

where H is the hop size in samples

From the right plot in Fig. 5, it can be seen that by applying the proposed method, periodicity T2 is not comprised within regions $R2$ (regions $R1$ are outside the displayed lag range). Thus, the song's beat period is estimated at the periodicity corresponding to MD, which corresponds to the beat prominence $B1$.

IV. Results

Details of the experimental framework utilized in order to evaluate the robustness of the proposed tempo detection algorithms are given in this section. First, a short description of the three different databases of songs is given. Following this, the metrics and statistical test utilized to evaluate the presented tempo detection method are introduced. Next, the different configurations for evaluation of the proposed tempo detection are detailed. Finally, the results of the evaluation are presented.

A. Databases

Three different databases of music signals and the corresponding manually annotated tempos are utilized in order to evaluate the robustness of the algorithm. The first database, which is denoted as $Db1$, was used by Klapuri *et al.* in the meter detection system presented in [11]. The other two databases, denoted as $Db2$ and $Db3$, were made publicly available by the organizers of the tempo detection contest in ISMIR 2004. The

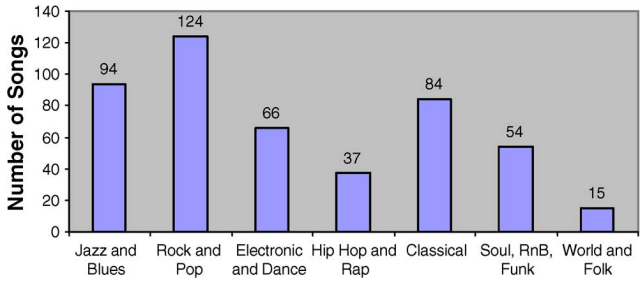

Fig. 6. Details of the database of song excerpts used in [11].

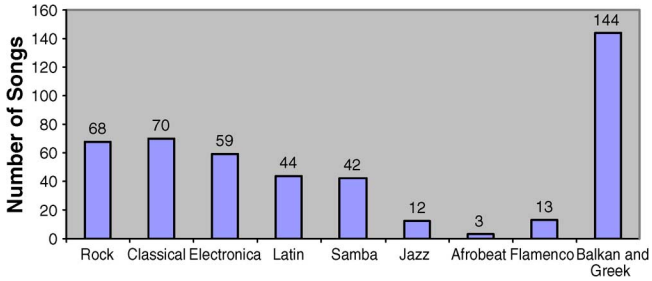

Fig. 7. Genre distribution of the song excerpts database [19].

results of the contest are published in [19]. These databases have also been used to evaluate more recent tempo detection methods [6], [13], [14]. Details of the three databases are given as follows.

- ***Db1***: *song excerpts database used in* [11].
  The manually annotated database used by Klapuri in [11] does not include tempo information. However, in the presented evaluation the tempo of each song is annotated by calculating the median of the manually annotated beat times used in [11]. The database comprises excerpts of 474 different songs of an approximate duration of 1 m. A summary of the distribution of the 474 songs according to the music genre is shown in Fig. 6. A more detailed description of the database is given in [26].
- ***Db2***: *database of song excerpts used in ISMIR 04 tempo detection contest* [19]
  This database is comprised of 465 song excerpts, which have a duration of approximately 20 s each. A summary of the genre distribution of the database is shown in Fig. 7.
- ***Db3***: *database of ballroom songs used in ISMIR 04 tempo detection contest* [19]
  This database is comprised of 698 excerpts of ballroom music, which have an approximate duration of 30 s each. A summary of the genre distribution of the database is shown in Fig. 8.

B. Evaluation Metrics

The metrics applied in the tempo detection contest in ISMIR 2004 are used here in order to evaluate the proposed tempo detection system [27].

- ***Acc1***: a correct tempo detection estimate will fall within a 4% window of the ground-truth tempo
- ***Acc2***: a correct tempo detection estimate will fall within a 4% window of either the ground-truth tempo, or half, double, triple, or one third of the ground truth tempo

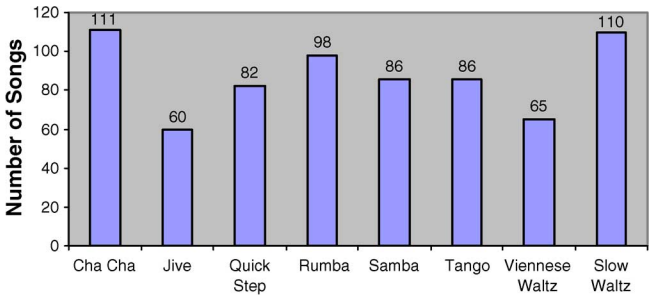

Fig. 8. Style distribution of the ballroom dance music excerpts [19].

Both MIREX 2005 and 2006 tempo detection evaluations use a collection of perceptual tempo annotations [21], [28], [29]. The aim of the perceptual ground truth was to identify the two most perceptually salient tempi in a piece of music. This was achieved by annotating a collection of 160 excerpts using 40 annotators per song. A ground truth for each excerpt is derived from the two highest peaks in the perceptual tempo distribution and their relative salience. Then, each tempo-extraction algorithm generates two tempo values for each musical excerpt and its performance is measured by its ability to match the two ground-truth tempi. The database of songs used in MIREX 2005 and 2006 tempo detection evaluations and its corresponding ground truths are not publicly available. Consequently, in the results presented here, $Acc1$ and $Acc2$ metrics are used. This allows the methods presented here to be compared against the results presented in ISMIR 2004 tempo detection contest [19]. This also facilitates comparisons against evaluations that used the same databases and metrics [13], [14].

C. Statistical Significance

In order to ensure that results obtained by the above metrics are statistically valid, the significance of the difference in performance of the tempo detection methods should be estimated. The error rates of the presented tempo detection systems are analyzed using McNemar's tests, which are used to determine statistical significance when comparing the performance of system pairs [30]. McNemar's test has a low probability of incorrectly detecting a difference when no difference exists as well as good discriminative power (the ability to detect a difference where one does exist) [30].

McNemar's test returns a P value when comparing the performance of two systems. If P is less than the threshold, the difference is considered "statistically significant." If the P value is greater than the threshold the difference is "not statistically significant" and both systems are considered to perform similarly for the data given. Dietterich sets the threshold at which the result be considered significant to $P < 0.05$, an arbitrary value that has been widely accepted in method performance evaluations. However, other research has also used $P < 0.01$ as an alternative threshold [19]. In the results presented in this paper, $P < 0.05$ is used as a threshold of significance. However, in order to allow further interpretation of the results presented, actual P values are reported.

TABLE II
MULTIBAND CONFIGURATIONS

	Bands		
Configuration name	**Low**	**Middle**	**High**
SCa	*SC*	*SC*	*SC*
Hyb1	*SC*	*SC*	*TD*
Hyb2	*SC*	*TD*	*TD*

D. Proposed Tempo Detection Approach Configurations

As described in Section III, the proposed tempo detection system allows for different configurations. The multiband configurations used in the evaluation can be seen in Table II, where SC and TD denote the use of the spectral complex change onset detector and the transient detector, respectively [23], [25]. As can be seen from the table, the first configuration uses the same onset detector SC in each band. The remaining two configurations use the hybrid multiband approaches shown in Table I.

The proposed tempo detection method is based on Davies *et al.*'s approach [7]. Thus, in order to track the improvement of the proposed approach from Davies *et al.*'s model, results obtained by using our implementation of Davies *et al.*'s model are also included as a reference. This implementation is denoted as $Davies2$,[6] which can be interpreted as the model depicted in Fig. 2 without the multiband decomposition, with only the use of the complex change onset detector, and where both the original weighting method and comb filter method used in [7] are applied (see (10) and (7) respectively). In addition, the use of TD within $Davies2$ model instead of SC is also evaluated using the entire dataset. This single band model is denoted as $Davies2 + TD$, and it is used to investigate the impact of using TD within $Davies2$ model.

E. TEST1: Evaluation of the Suggested Tempo Detection Configurations

In this section, the suggested tempo detection approaches are evaluated. In addition, the weighting and comb filter methods introduced in Section III-E and III-D, respectively, are evaluated, which are a modification of the techniques used in Davies *et al.*'s tempo detector [7]. In order to investigate the impact of using the proposed modifications, the following evaluations are performed. First, the tempo detection methods are evaluated using the original comb filter method [see (7)] and the weighting method introduced in [7]. Following this, the first modification is applied by using the proposed weighting method (see Section III-E). Next, the second modification is applied by using the proposed comb filter method [see (8)].

Evaluation of Weighting Methods: The advantage of using the proposed weighting method is evaluated as follows; first, the tempo detection methods are evaluated for the entire dataset using the original configuration of comb filter-weighting methods used in [7]. Then, this original configuration is modified by using the proposed weighting method introduced in Section III-E. The results of this evaluation are shown in Table III. In addition, P values obtained using McNemar's

[6]Since this model is our implantation of [7] and not the original implementation used by the authors of [7], our implementation of that model is denoted as "$Davies2$".

TABLE III
TEMPO ACCURACY RESULTS BY APPLYING DIFFERENT WEIGHTING METHODS FOR THE ENTIRE DATASET. [7]'S COMB FILTER METHOD IS USED TO GENERATE THE RESULTS. P VALUES ARE GENERATED USING MCNEMAR'S TESTS

	Acc1			*Acc2*		
Weighting method	*Original*	*Proposed*	$P_{(original\text{-}proposed)}$	*Original*	*Proposed*	$P_{(original\text{-}proposed)}$
SCa	63.6 %	**64.5 %**	**0.0180**	89.7 %	**90.6 %**	**0.0180**
Hyb1	63.4 %	**64.9 %**	**< 0.0001**	89.5 %	**91.4 %**	**< 0.0001**
Hyb2	62.5 %	**63.8 %**	**0.0005**	89.1 %	**91 %**	**< 0.0001**
Davies2	62.3 %	**63.8 %**	**0.003**	88.3 %	**89.5 %**	**0.0042**
Davies2+TD	59.6 %	**61.8 %**	**<0.0001**	86.9 %	**90 %**	**<0.0001**

TABLE IV
TEMPO ACCURACY RESULTS BY APPLYING DIFFERENT COMB FILTER METHODS FOR THE ENTIRE DATASET. THE SUGGESTED WEIGHTING METHOD IS USED TO GENERATE THE RESULTS. P VALUES ARE GENERATED USING MCNEMAR'S TESTS

	Acc1			*Acc2*		
Comb filter method	*Original*	*Proposed*	$P_{(original\text{-}proposed)}$	*Original*	*Proposed*	$P_{(original\text{-}proposed)}$
SCa	64.5 %	**65.3 %**	0.2188	90.6 %	**91.6 %**	**0.0489**
Hyb1	**64.9 %**	64.7 %	0.8989	91.4 %	**92.7 %**	**0.0034**
Hyb2	**63.8 %**	63.6 %	0.8149	91 %	**92.1 %**	**0.0147**
Davies2	**63.8 %**	62.8 %	**0.0472**	**89.5 %**	88.5 %	0.0648
Davies2+TD	**61.8 %**	61.5 %	0.6350	**90 %**	89.3 %	0.2077

Tests in order to compare the performance of each tempo detection model using both weighting methods are also shown in Table III.

The result obtained by the best weighting method for each tempo detection model is highlighted in bold in Table III, which shows that the use of the proposed weighting method improves the results for all tempo detection models in both $Acc1$ and $Acc2$ metrics. As can be seen in the table, all method comparisons provide statistically significant results.

Evaluation of Comb Filter Methods: The periodicity detection method used in the proposed tempo detection models is a modification of the method used in Davies *et al.* tempo detection method [7]. Consequently, the advantage of the use of the weighted comb filter method introduced in (8) is evaluated in this section. In order to investigate the difference in performance from the previous evaluation (see Table III), the suggested weighting method was used within the evaluated tempo detection methods. The results of this evaluation are shown in Table IV. In addition, P values obtained using McNemar's Tests in order to compare the performance of each tempo detection model using both comb filter methods are also shown in Table IV.

From Table IV, the use of the original comb filter method performs better than the use of the proposed comb filter for the single-band methods $Davies2$ and $Davies2 + TD$ using both metrics. However, the proposed comb filter method is a better weighting method using the $Acc2$ metric for the proposed multiband detection models, which P values also show significant performance differences. This might due to the characteristics of $Davies2$ model, which only uses a single $PeDF$. Thus, the original comb filter method captures better the periodicities of different metrical levels using a nonweighted comb filter template. However, the multiband methods combine different band $PeDFs$, which might represent individually the periodicities of different metrical levels.

TABLE V
MCNEMAR TEST COMPARISONS (P VALUES) OF TEMPO DETECTION METHODS. THE LOW AND HIGH SIDE OF THE MAIN DIAGONAL OF THE TABLE CORRESPOND TO P VALUES OBTAINED USING $Acc1$ AND $Acc2$ METRIC COMPARISONS, RESPECTIVELY

METHOD	*SCa*	*Hyb1*	*Hyb2*	*Davies2*	*Davies2+ TD*
SCa		**0.0472**	0.52	**<0.0001**	**<0.0001**
Hyb1	0.3799		0.1649	**<0.0001**	**<0.0001**
Hyb2	**0.0215**	**0.0377**		**<0.0001**	**<0.0001**
Davies2	**<0.0001**	**0.0012**	0.1393		0.0734
Davies2+TD	**<0.0001**	**<0.0001**	**<0.0001**	**0.0021**	

Statistical Differences Between the Tempo Detection Methods: P values obtained by comparing the performance of the evaluated tempo detection methods are shown in Table V using the entire dataset. In the comparisons, the multiband methods used the proposed modifications. The low and high side of the main diagonal of Table V correspond to comparisons obtained using $Acc1$ and $Acc2$ metrics, respectively. As an example, $Hyb1 - SCa = 0.3799$ (low side) and $Hyb1 - SCa = 0.0472$ (high side) correspond to P values obtained using $Acc1$ and $Acc2$, respectively. Statistically significant comparisons in Table V are highlighted in bold.

From Table IV, the best result for $Acc1$ is obtained using SCa (65.3%, highlighted and underlined). However, SCa does not show a statistical performance difference from $Hyb1$ in Table V ($P = 0.3799$) and both method performances can be interpreted as being similar using $Acc1$. By considering $Acc2$, the accuracy obtained by $Hyb1$ is equal to 92.7% (highlighted and underlined). However, as can be seen in Table V, $Hyb1$ does not show a statistical difference from $Hyb2$ ($P = 0.1649$). More explicitly, $Hyb1$ can be seen as the best overall method, which

TABLE VI
TEMPO ACCURACY RESULTS USING BETWEEN EXISTING APPROACHES AND THE PROPOSED APPROACHES FOR $Db1$, $Db2$ AND $Db3$ [8]

	Db1		*Db2*		*Db3*	
METHOD	*Acc1* (%)	*Acc2* (%)	*Acc1* (%)	*Acc2* (%)	*Acc1* (%)	*Acc2* (%)
Sca	**74.5**	91.8	50.4	88.8	68.9	93.4
Hyb1	73.2	91.6	49.4	91.6	**69.2**	**94.1**
Hyb2	71.3	**92.2**	48.9	**91.8**	68.1	92.1
Davies2	73.8	88.6	46.8	82.3	64.8	92
Eck			**60**	79	63	91[9]
Klapuri			58.49	91.18	63.18	90.97
Ellis			45.8	80.6		

performance is similar to SCa and $Hyb2$ using $Acc1$ and $Acc2$, respectively.

As can be seen in Table IV, methods that depend heavily on the accuracy of SC ($Davies2$ and SCa) perform better for $Acc1$ metric. On the other hand, the performance of methods that depend more on the accuracy of TD ($Hyb2$ and $Davies2 + TD$) improve the performance using $Acc2$ metric.[7] This can also be seen by comparing SCa and $Hyb1$, where the use of TD in the HFB provides a statistically significant improvement in $Acc2$ ($P = 0.0472$ in Table V).

F. TEST2: Comparison of the Suggested Tempo Detection Methods Against Other Existing Approaches

In order to evaluate in more detail the accuracy of the different multiband configurations introduced in Section IV-C, the proposed multiband approaches and $Davies2$ system using its original weighting method are evaluated for the three databases $Db1$, $Db2$ and $Db3$.

The results for $Db1$ are shown in Table VI, where the results obtained by the proposed approaches and especially by $Hyb2$ using $Acc2$ (92.2%, highlighted in bold) improve the results obtained by $Davies2$ (88.6%). It can be seen that methods using SC (SCa, $Hyb1$ and $Davies2$) improve upon $Hyb2$ using $Acc1$.

The results obtained by the evaluated methods in $Db1$ are sorted by genre and displayed in Fig. 9 for the accuracy metric $Acc2$. The results show that the difference in performance between classical music and the other genres is remarkable, where the best performance for classical music attains an accuracy of 72.6% using $Hyb2$. In contrast, the minimum accuracy for the other genres using the multiband approaches is equal to 93.3%.

The results obtained by evaluating the presented approach using $Db2$ and $Db3$ are shown in Table V. As a reference, the results obtained by the winner of the ISMIR tempo detection contest A. Klapuri are also included in the comparison [27]. In addition, the evaluation presented by Eck's model in [13] using the same databases, $Db2$ and $Db3$, is also included. In [14], Ellis uses $Db2$ to evaluate his approach; the results by his method are also included in the comparison.

Considering $Db2$ results, it can be seen that the proposed $Hyb2$ model obtains an $Acc2$ accuracy of 91.8% (highlighted in underline), which is only slightly better than Klapuri (91.18%) and significantly better than $Davies2$ (82.3%), Eck (79%) and Ellis (80.6%). However, by using $Acc1$, results reported by existing approaches. Eck (60%) and Klapuri (58.49%), improve over the proposed methods.

Considering $Db3$ results, the accuracy using metric $Acc2$ obtained by the proposed $Hyb1$ model is equal to 94.1%, which improves upon the results provided by $Davies2$ (92%), *Klapuri* (90.97%), and *Eck* (91%). The results obtained by $Hyb1$ using $Acc1$ (69.2%) also clearly improve existing models, where the best performing method was $Davies2$ (64.8%).

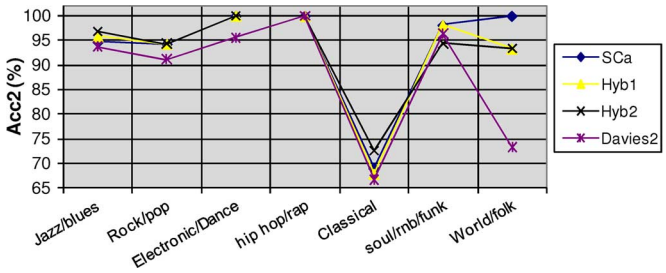

Fig. 9. Genre distribution of tempo accuracy.

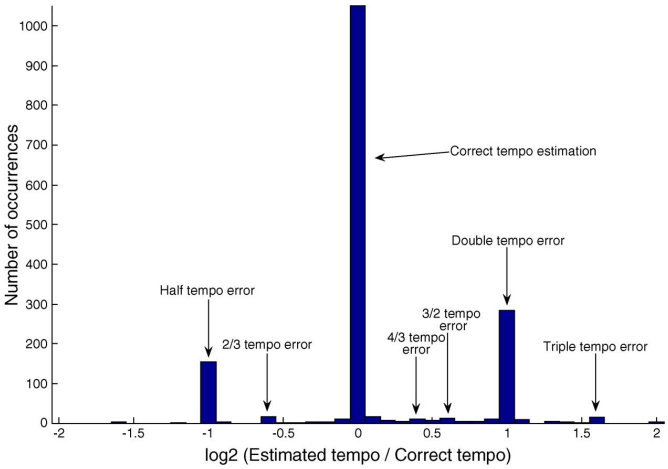

Fig. 10. Histogram of ratio between estimated tempo and ground truth tempo for entire data set using $Hyb1$.

G. Tempo Error Analysis of Hyb1 Method

As mentioned in Section IV-E, $Hyb1$ is the best overall method for the entire dataset. The error analysis of tempo estimates using $Hyb1$ is analyzed in this section. A histogram of the ratio between the estimated tempi and the manually annotated ground truth is shown in Fig. 10. It can be seen that, as expected, the most occurring errors correspond to double and half errors. Small peaks in the histogram corresponding to less frequently occurring errors can also be seen. As an example, in the case of songs with a crotchet as the primary metrical level, a 2/3 or 4/3 error ratio typically arises from dotted crotchet and dotted quaver periodicities, respectively. In contrast, a 3/2 error ratio commonly corresponds to periodicities resulting from compound metre subdivisions.

In Fig. 11, the error distribution of $Hyb1$ method with respect to the ground truth annotations and error ratio type is displayed, where it can be seen that $Hyb1$ method does not estimate correctly any song played at a tempo lower than 70 bpm or greater

[7]The difference in performance between metrics $Acc1$ and $Acc2$ is discussed further in Section V.

[8] Published evaluations that do not include results using $Db1$ or $Db3$ are grayed out in the table.

[9] No decimals were used in the results presented by Eck *et al.* in [13].

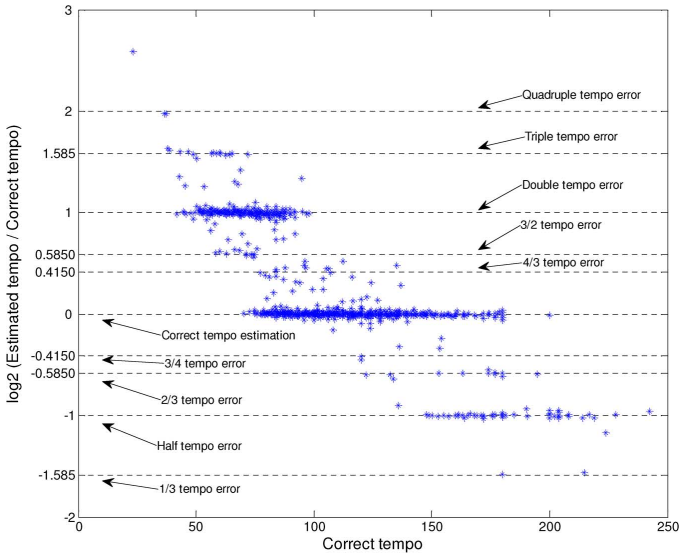

Fig. 11. Error distribution of $Hyb1$ with respect to tempi and error type for entire data set.

than 200 bpm. As can be seen in the figure, the density of double and half tempo errors increase remarkably with tempi lower than 90 bpm and higher than 150 bpm, respectively.

V. Discussion

In the previous section, an evaluation of the presented tempo detection models was presented. In this section, the following topics are given further discussion.

A. Modifications Applied to $Davies2$ Reference Model (Weighting Method, Comb Filter Method, and Multiband Configuration)

The proposed tempo detection approach is based on the model presented in [7]. A set of modifications to this model, which is denoted as $Davies2$, have been suggested in Section III and evaluated in the previous section. First, the proposed weighting method is evaluated in Table III, where it can be seen that replacing the original weighting method in $Davies2$ model, improves $Davies2$ model from 62.3% to 63.8% using $Acc1$, and from 88.3% to 89.5% using $Acc2$ metric for the entire data set. As can be seen in Table III, the use of the proposed weighting method provides statistically significant performance differences in all the methods that took part of the evaluation.

The second modification uses the suggested weighted comb filter in the $PeDFs$ of the multiband approaches. The use of the weighted comb filter provides performance differences in the multiband approaches but not in the $Davies2$ model, where the use of its original comb filter is a more suitable method. Since $Davies2$ generates a single $PeDF$, the metrically unbiased comb filter used in [7] captures better existing periodicities in higher metrical levels. In contrast, the use of the weighted comb filter proves to be more accurate within a multiband decomposition, which shows statistically performance differences using $Acc2$ metric.

The third modification consists of turning $Davies2$ into a multiband configuration, which is denoted as SCa. By considering Table III, it can be seen that results improve further, where the best results for SCa correspond to 65.3% and 91.6% using $Acc1$ and $Acc2$, respectively. The fourth modification, which consists on using a hybrid model, is discussed in the following section.

B. Hybrid Multiband Tempo Detector Comparison

Results are generally improved over the reference model $Davies2$ by using the proposed hybrid multiband configurations $Hyb1$ and $Hyb2$. From Table III and Table IV, $Hyb1$ can be seen as the best overall method, in which the performance is statistically similar to SCa and $Hyb2$ using $Acc1$ and $Acc2$, respectively. The only difference between $Hyb1$ and $Hyb2$ models lies in the middle frequency band, which uses the SC and TD onset detectors, respectively. The SC performs better than the TD in tracking changes resulting from solos or slow onset instrument in soft melodies. This explains the reason why $Hyb1$ provides better results than $Hyb2$ in the ballroom database $Db3$.

Differences in performance between the hybrid methods were also found by using $Acc1$ and $Acc2$ metrics. $Hyb1$ was generally a better method using $Acc1$. However, the performance of $Hyb2$ improves using $Acc2$, which uses TD in both bands MFB and HFB. TD is not an energy dependant method and does not necessarily generate more prominent peaks in the ODF when notes on the beat are played more accented. Thus, percussive events located at beat subdivisions will generate equally prominent peaks. This can lead to tempo estimates in multiples of the perceived tempo, which will be estimated as a correct estimation using $Acc2$ but not if metric $Acc1$ is used instead. The same principle can also be seen in both SCa and $Davies2$, which in relation to the hybrid methods perform better for $Acc1$ metric.

C. Comparison Against State of the Art Tempo Detectors

Databases $Db2$ and $Db3$ allow comparisons against existing published research. The best results for the evaluated methods using $Acc2$ correspond to 91.8% ($Hyb2$) and 94.1% ($Hyb1$) using databases $Db2$ and $Db3$, respectively. Both hybrid multiband configurations compare favorably against other existing approaches, where the best results for $Db2$ and $Db3$ correspond to 91.18% (Klapuri) and 92% ($Davies2$), respectively. The results for the metric $Acc1$ are also included in the evaluation, where the best methods for $Db2$ and $Db3$ correspond to *Eck* (60%) and. $Hyb1$ (70.2%), respectively. The only test in which $Hyb1$ and $Hyb2$ did not compare favorably upon existing approaches is using metric $Acc1$ for $Db2$. In this case, both *Eck* and *Klapuri* significantly improve upon the hybrid configurations.[10] Klapuri utilized the same algorithm to win the last tempo detection context organized in MIREX 2006, which makes the presented results more significant. However, it should be noted that *Klapuri* had the disadvantage of being the only algorithm in the comparison presented in the previous section that did not have access to the databases prior to the evaluation. In [6], Peeters uses $Db3$ in order to evaluate his tempo detection model, which obtained an accuracy of 92%. However, different metrics were used to evaluate the model, and therefore it is not directly compared against the methods presented in Section IV. In [31],

[10]Hybrid methods $Hyb1$ and $Hyb2$ only obtained 49.4% and 48.9%, respectively, for $Db2$ using metric $Acc1$.

a recent model is reported, which uses pre-knowledge of the rhythmic style that comprise $Db3$ (see Fig. 8). This classification based algorithm, which is evaluated using 90% and 10% of data in the training and testing phases, respectively, provides results equal to 85.8% and 94.4% using $Acc1$ and $Acc2$, respectively, for $Db3$.

D. Performance Metric

The metric $Acc2$ accounts for the deviation factors occurring in tempo perception described in Section I [11], [13], [19]. Consequently, $Acc2$ might appear as a more suitable metric to evaluate a tempo detector than $Acc1$, which might be biased towards the metrical perception of the database annotator. It is interesting to note that Eck *et al.*'s model provided the best and worse results for database $Db2$ using the metrics $Acc1$ and $Acc2$, respectively. This might indicate that $Acc1$ measures the ability of the model to match the metrical perception of the annotator, which could vary if a different annotator is used. However, this depends heavily on the music style being analyzed; as an example, $Db3$ comprises ballroom dance music excerpts, which are composed to be danced at a given tempo. Thus, the $Acc1$ is a suitable metric for database $Db3$.

However, the $Acc2$ metric does not take meter into account. A more adequate tempo metric is used in [6], which considers an estimated tempo as "correct" if factors 2 or 1/2 occur for duple meter music, or factors 3 or 1/3 occur for both compound meters and triple meters. The perceptual metrics introduced in MIREX 2005 and 2006 implicitly accounts for the meter of the music [21], [28], [29]. Since a large number of humans are used to generate the perceptual distribution of tempo of each song, it is very likely that only "correct" metrical deviations will be selected as the two ground truth tempi.

VI. Conclusion and Future Work

A novel tempo detection system has been presented in this paper. First, a literature review of existing research in the area is given in Section II, which includes a number of research avenues that can lead to potential improvement upon the accuracy of existing methods. The proposed system is introduced in Section III, which suggests the application of a set of modifications to Davies *et al.* model [7]. The modifications include the following.

- The use of an improved weighting method, which improves the results in all tempo detection methods that took part of the evaluation.
- The use of a weighted comb filter method, which improves the results in all multiband tempo detection methods.
- The use of a multiband decomposition; It was shown that adapting Davies *et al.* model to a multiband configuration improves the results. In addition, hybrid multiband configurations which combine the use of unique onset detectors for each frequency band were also introduced. This modification also improved further the results.

These contributions were evaluated in Section IV, where the presented system compared favorably against existing approaches for three different databases. A discussion of potential avenues of future work is described as follows.

- By considering Fig. 9, it is apparent that a robust method capable of detecting the tempo in classical music is yet to be implemented, which suggests that further research in the area is still required.
- The results presented in Section IV show that the choice of the *onset detector* has a significant impact on the accuracy. It was shown that the use of a hybrid multiband configuration that uses different types of onset detector improves the results. However, different hybrid configurations provide different results for $Db3$ and the other databases. Consequently, a system that dynamically chooses the most appropriate onset detector in each band should be implemented. This might be achievable by detecting transients in both MFB and HFB. Thus, the TD will be used in each tempo detector band only if a certain number of transients are detected in the band.
- The proposed approach uses the autocorrelation function to generate the *periodicity detection* function. Since only the main periodicity needs to be extracted, the autocorrelation function provides sufficient accuracy. By informal testing, no major difference was noticed between alternative periodicity detection models. The presented model captures periodicities in *higher metrical levels* by applying a comb filter to the autocorrelation function. It was noticed that different comb filters improve the performance of a single-band or a multiband model independently. However, in contrast to [9] and [11], lower metrical information such as the tatum was not used. The advantage of using such information within the proposed system warrants future work.
- The advantage of using the proposed weighting method was evaluated in Section IV. However, as can be seen in Fig. 11, the proposed tempo detection model has difficulties to track slow and very fast tempi, which can be a result of the weighting function used. This clearly requires further investigation.
- It is also shown that the use of a *multiband decomposition* is important. The proposed system uses three frequency bands, in which cutoff frequencies are chosen to cover regions of common activity for certain instrument types. In the model, each band contributes equally to the overall periodicity estimation. A more dynamic multiband decomposition should be envisaged. Thus, the reliability of the extracted periodicities in each individual band will be evaluated. This ensures that only bands in which onset detection functions provide valuable periodicities will be used. As an example, a song with no presence of low-frequency instruments should not have a specific LFB in the tempo detection model. Another potential improvement of the multiband model may be the use of a multiresolution approach, as suggested in [32] for onset detection purposes. Thus, lower bands where onsets take more time to reach the maximum of the onset amplitude can use longer frames in the time–frequency analysis than in high bands. In contrast, in order to improve the system time resolution, sharp transients can be tracked in high bands by using short frames.

ACKNOWLEDGMENT

The authors would like to thank D. Barry, D. Fitzgerald, and D. Dorran for proofreading the paper. The authors would also like to thank A. Klapuri for allowing access to the database of signals used in [11].

REFERENCES

[1] C. Drake, L. Gros, and A. Penel, "How fast is that music? The relation between physical and perceived tempo," in *Proc. Int. Conf. Music Percept. Cognit.*, 1999, Seoul National Univ..

[2] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Comput. Music J.*, vol. 29, pp. 34–54, 2005.

[3] M. F. McKinney and D. Moelants, "Deviations from the resonance theory of tempo induction," in *Proc. Conf. Interdisciplinary Musicol.*, 2004.

[4] M. F. McKinney and D. Moelants, "Ambiguity in tempo perception: What draws listeners to different metrical levels?," *Music Percept.*, vol. 24, pp. 155–166, 2006.

[5] J. Laroche, "Estimating tempo, swing and beat locations in audio recordings," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2001, pp. 135–138.

[6] G. Peeters, "Time variable tempo detection and beat marking," in *Proc. Int. Comput. Music Conf., ICMC*, Barcelona, Spain, 2005, p. 0.6.

[7] M. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1009–1020, Mar. 2007.

[8] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of musical signals," in *Proc. 5th Int. Symp. Music Inf. Retrieval (ISMIR)*, 2004.

[9] C. Uhle, J. Rohden, M. Cremer, and J. Herre, "Low complexity musical meter estimation from polyphonic music," in *Proc. Audio Eng. Soc. 25th Int. Conf.*, 2004, pp. 63–68.

[10] S. Dixon, E. Pampalk, and G. Widmer, "Classification of dance music by periodicity patterns," in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2003, pp. 159–165.

[11] A. Klapuri, A. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 342–355, Jan. 2004.

[12] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, pp. 588–601, 1998.

[13] D. Eck and N. Casagrande, "Finding meter in music using an autocorrelation phase matrix and Shannon entropy," in *Proc. ISMIR*, 2005, vol. 300, pp. 350–400.

[14] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res., Special Iss. Beat and Tempo Extraction*, vol. 36, pp. 51–60, 2007.

[15] M. Alonso, B. David, and G. Richard, "Tempo extraction for audio recordings," in *Proc. 1st Annu. Music Inf. Retrieval Eval. eXchange, MIREX*, 2005.

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

[17] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proc. IEEE Int. Conf. Multimedia Expo, ICME*, 2001, pp. 881–884.

[18] A. Pikrakis, I. Antonopoulos, and S. Theodoridis, "Music meter and tempo tracking from raw polyphonic audio," in *Proc. 5th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2004.

[19] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1832–1844, Sep. 2006.

[20] F. Gouyon and B. Meudic, "Towards rhythmic content processing of musical signals: Fostering complementary approaches," *J. New Music Res.*, vol. 32, pp. 41–64, 2003.

[21] M. F. McKinney, D. Moelants, M. E. P. Davies, A. Klapuri, and U. K. London, "Evaluation of audio beat tracking and music tempo extraction algorithms," *J. New Music Res.*, vol. 36, pp. 1–16, 2007.

[22] M. Davies and M. D. Plumbley, "Comparing mid-level representations for audio based beat tracking," in *Proc. DMRN Summer Conf.*, Glasgow, U.K., 2005.

[23] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. 6th Int. Conf. Digital Audio Effects (DAFx-03)*, London, U.K., 2003.

[24] F. Gouyon, S. Dixon, G. Widmer, and I. Porto, "Evaluating low-level features for beat classification and tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. 1309–1312.

[25] D. Barry, D. Fitzgerald, E. Coyle, and B. Lawlor, "Drum source separation using percussive feature detection and spectral modulation," in *Proc. Irish Signals Syst. Conf., ISSC*, Dublin, Ireland, 2005.

[26] A. Klapuri, 2008, Details of the evaluation database used in "Analysis of the meter of acoustic musical signals" [Online]. Available: http://www.cs.tut.fi/~klap/iiro/meter/database.html Accessed [14th Aug. 2009]

[27] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1832–1844, Sep. 2006.

[28] MIREX, 2005, "Audio Tempo Extraction Contest," [Online]. Available: http://www.music-ir.org/mirex/2006/index.php/Audio_Tempo_Extraction Accessed [14th Aug. 2009]

[29] MIREX, 2006, "Audio Tempo Extraction Contest," [Online]. Available: http://www.music-ir.org/mirex/2006/index.php/Audio_Tempo_Extraction Accessed [14th Aug. 2009]

[30] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, pp. 1895–1923, 1998.

[31] M. E. P. Davies and M. D. Plumbley, "Exploring the effect of rhythmic style classification on automatic tempo estimation," in *Proc. 16th Eur. Signal Process. Conf., EUSIPCO*, 2008.

[32] C. Duxbury, J. P. Bello, M. Sandler, and M. Davies, "A comparison between fixed and multiresolution analysis for onset detection in musical signals," in *Proc. 7th Int. Conf. Digital Audio Effects (DAFx)*, Naples, Italy, 2004.

Mikel Gainza received the Ph.D. degree in digital signal processing from the Dublin Institute of Technology (DIT), Dublin, Ireland, in 2006, specializing in audio content analysis.

He is a Researcher and Lecturer in the Audio Research Group, DIT. He has contributed to several major research projects in the field of music information retrieval and audio content analysis, which includes both European Framework and nationally funded research projects. He is also involved in the supervision of several Ph.D. students in the Audio Research Group. His research interests include music information retrieval systems, audio content analysis (music and speech), rhythm description, automatic music transcription, and music summarization.

Eugene Coyle is the Head of the School of Electrical Engineering Systems , Dublin Institute of Technology, Dublin, Ireland, and advisor to the School's Audio Research Group. His research spans the fields of control, electrical, digital signal processing, and biomedical and rehabilitation engineering. He has participated in numerous research projects over the past 20 years as both Researcher and Principal Investigator.

Prof. Coyle is a Chartered Engineer and Fellow of the Institution of Engineering and Technology (IET), Engineers Ireland (IEI), and the Energy Institute (EI). He is currently the chairperson of IET Ireland Network, and he is also a member by invitation of the Engineering Advisory Committee to the Frontiers Engineering and Science Directorate of Science Foundation Ireland, SFI.