Modeling the perception of tempo

Anders Elowsson and Anders Friberg
School of Computer Science and Communication, Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

(Received 22 July 2014; revised 27 March 2015; accepted 4 April 2015)

A system is proposed in which rhythmic representations are used to model the perception of tempo in music. The system can be understood as a five-layered model, where representations are transformed into higher-level abstractions in each layer. First, source separation is applied (Audio Level), onsets are detected (Onset Level), and interonset relationships are analyzed (Interonset Level). Then, several high-level representations of rhythm are computed (Rhythm Level). The periodicity of the music is modeled by the cepstrum vector—the periodicity of an interonset interval (IOI)–histogram. The pulse strength for plausible beat length candidates is defined by computing the magnitudes in different IOI histograms. The speed of the music is modeled as a continuous function on the basis of the idea that such a function corresponds to the underlying perceptual phenomena, and it seems to effectively reduce octave errors. By combining the rhythmic representations in a logistic regression framework, the tempo of the music is finally computed (Tempo Level). The results are the highest reported in a formal benchmarking test (2006–2013), with a P-Score of 0.857. Furthermore, the highest results so far are reported for two widely adopted test sets, with an $\text{Acc}_1$ of 77.3% and 93.0% for the Songs and Ballroom datasets.

© 2015 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4919306]

I. INTRODUCTION

The rhythmic structure of music often induces the perception of pulse sensations at different time scales (Eronen and Klapuri, 2010). As this is an emergent perceptual property of the musical sound, it can only be defined in terms of listening experiments or similar estimations of perceptual aspects. Pulse and rhythm in music is closely related to human gestures and dancing. Thus, a common definition of the main pulse, here referred to as the tactus or beat level, is the pulse scale that a listener would tap in conjunction with the music using their foot or hand. The length between two consecutive taps will be referred to as the beat length (London, 2012). The rate of the tactus pulse scale corresponds to the tempo of the music and is expressed as the number of beats per minute (BPM). An alternative definition of pulse (and a source of confusion) is the notated meter in the score. The meter corresponds to how the composer envisions that the listener or musician should hear the pulse and it is thus indicating how to perform the piece in this aspect. However, the notated pulse may also reflect some notational conventions, such that a particular style is often notated in a certain way. In many cases, the notated pulse and the perceptual pulse will coincide. However, different listeners may hear the main perceptual pulse as occurring on different temporal levels. This may reflect an ambiguity of the different tactus levels that emerge from the musical surface in terms of the rhythmic patterns. Therefore, it is more relevant to allow different perceptual tactus levels each associated with a salience or weight factor.

Tempo information is an important parameter in music. It is useful to many different applications such as automated DJ mixing as well as mood or genre identification (Gouyon et al., 2006). Another field where there is a potential need for tempo estimation from music audio is automated music mixing; it is common practice to use delay and compressor release settings that are in synchronization with the tempo of the music.

A difficult problem when estimating tempo is to find the correct tempo octave (i.e., to find the pulse sensation that correspond to the tactus level and not the tatum or measure level). Erroneous estimations are often produced at half or double the correct tempo, as well as one third or three times the tempo for music with a triple meter structure. One aspect of music that is essential when locating the correct tempo octave is the perceived speed of the music, as speed and tempo are correlated (Madison and Paulin, 2010; Levy, 2011). Hockman and Fujinaga (2010) have presented a classifier of music as fast or slow, and in two recent papers (Elowsson and Friberg, 2013a; Elowsson et al., 2013), the speed of music has been modeled as a linear function of onset densities, tempo and spectral fluctuations. It seems reasonable to assume that a model similar to this (excluding any prior tempo estimation) could be used to overcome the octave error in tempo estimation.

A. Previous work

Numerous systems that try to estimate tempo of music audio have been proposed in recent years. A common denominator among the systems is to start with a measure of change in energy over time. Energy fluctuations are often tracked by using filterbanks distributed on the mel-scale (see, e.g., Gkiokas et al., 2010). Similar subband processing...
for onset detection has been described by Dixon (2006) and Klapuri (1999) and can be motivated by its similarity to human hearing relating to critical bands (Duxbury et al., 2004). The periodicity at which the estimated energy changes over time is a key factor to detect tempo. Some systems detect periodicities directly by applying an autocorrelation function to the spectral flux (SF) curve (Alonso et al., 2004), while others (Seppänen, 2001; Gouyon et al., 2002) detect onsets in the music which are later analyzed in a histogram of interonset intervals (IOIs).

In recent years, a wide variety of approaches involving machine learning have been selected to predict the final tempo or tempo octave. One approach is to divide the songs into different classes such as slow, moderate and fast and then train a classifier to identify the different classes (Gkiokas et al., 2012; Levy, 2011). Chen et al. (2009) proposed a mood classifier to reassign the tempo octave of an estimation made by another algorithm and Xiao et al. (2008) used timbre information with the same purpose. It is striking that although the speed of music is a non-categorical continuous perceptual variable, it has never been modeled as a continuous function for tempo estimation.

Source separation has been proposed as a way to improve tempo estimation (Alonso et al., 2007; Chordia and Rae, 2009; Gkiokas et al., 2012). Source separation (harmonic/percussive) has also been shown to be useful to model the perceived speed of music (Elowsson et al., 2013), and it facilitates the detection of periodicities in percussive music. A recent trend (Seyerlehner et al., 2007; Eronen and Klapuri, 2010; Gkiokas et al., 2012) is to use the periodicity analysis of a database of musical excerpts (MEs) to generate templates for different tempi. These will be referred to as template-based methods. The tempo of the templates which are the most similar to the analyzed song is chosen as the estimated tempo. A somewhat similar approach is proposed by Peeters and Flocon-Chollet (2012), where a Gaussian mixture model is used to detect the relationship between four different audio features and perceptual tempo. Krebs and Widmer (2012) use a hidden Markov model (HMM) to locate beats and non-negative matrix factorization (NMF) to identify the tempo of the ME.

B. The multi-dimensionality of tempo

The variation in approaches may stem from the multi-dimensionality of the tempo estimation problem. The listener is presumably using speed related features such as onset densities in conjunction with pulse strengths at different metrical levels to infer the perceptual tempo. A perceptually motivated model should estimate tempo from data representations that incorporate these different high-level features of music. A benefit of using high-level abstractions is that they are generally invariant to local changes, so that they may entangle and hide different factors of variation in the data (Bengio et al., 2013). The dependency on accurate data representations has motivated research in deep belief networks (e.g., Lee et al., 2009), as well as research in perceptually defined features in Music Information Retrieval (MIR) (Friberg et al., 2014).

To find accurate data representations that cover different perceptual aspects of tempo is non-trivial and the majority of the proposed systems focus on one of them, or models them in distinctly separate steps. By training a method with templates (Seyerlehner et al., 2007; Gkiokas et al., 2012; Eronen and Klapuri, 2010) multiple dimensions are implicitly addressed. A problem with this approach may be its lack of generalization, i.e., the specific templates of the training set defines the tempo function instead of the general features of the training set such as onset densities. This can be remedied by using tempo classes (Gkiokas et al., 2012) to guide the final decision.

In this study, we will propose a different approach with a system called Tempo Estimation by Modeling Perceptual Speed (TEMPS) (Elowsson and Friberg, 2013b). One difference from previous approaches is that we estimate the speed of the music as a continuous function, to better model the underlying perceptual phenomena. We expect this approach to reduce the plausible errors at class boundaries of the classification methods (Gkiokas et al., 2012; Levy, 2011; Chen et al., 2009; Xiao et al., 2008). We also propose the cepstral vector—the periodicity of a periodicity function—as a way to compute a sparse and perceptually appropriate representation of the metrical structure. Finally we combine different measures of pulse strength to make the tempo estimation. We hope that the multi-layered structure that is used will make the system fairly invariant to local changes in the data.

II. PROBLEM DEFINITION

Tempo estimation is the task of finding the BPM in music. However, different listeners may hear the main perceptual pulse as occurring on different temporal levels. A number of evaluation metrics have been proposed to account for this octave error problem.

A. $\text{Acc}_1$ and $\text{Acc}_2$

A common approach is to use a single annotated tempo but report two measures of accuracy. The first one $\text{Acc}_1$ is used for correct estimations and the second one $\text{Acc}_2$ for estimations that are half or one third as well as double or three times the annotated tempo (e.g., Gouyon et al., 2006). We consider the $\text{Acc}_2$ measurement to be inappropriate, since in most cases only one or two of the allowed multiples are correct according to the perceptual salience of different tempo levels. This leads to perceptually erroneous estimations being counted as correct, which has also been noted by Gouyon et al. (2006). These errors represent one of the biggest challenges in tempo estimation and typical examples for $\frac{1}{4}$ meter (with an error at twice the tempo) and $\frac{3}{4}$ meter (with an error at three times the tempo) are shown in Fig. 1.

B. P-score

Another approach is currently used in the Audio Tempo Estimation task of the annual music information retrieval evaluation exchange (MIREX) (MIREX, 2013). In this competition, two annotated tempi ($\text{AT}_1$ and $\text{AT}_2$) are used as...
ground truth for each ME, based on the two most commonly tapped tempi among 40 listeners. The two annotations are assigned a weight between 0 and 1 according to their frequency of occurrence. Here, the weight of the most salient tempo AT1 is denoted $w$ with values 0.51 to 1, and the weight of AT2 will thus be $1-w$. Algorithms are requested to output two tempi ($T_1$ and $T_2$), and these are compared with the two ground truth annotations. The final score (P-Score) is the total weight of the ground truth annotations that are within 8% of any of the two tempo estimations. As noted by Gouyon et al. (2006) the width of the precision window is not a crucial factor. Returning to Fig. 1 it is easy to see the benefits with the P-Score measurement. Songs can have multiple tempi that are perceptually correct, and with the P-Score measurement the systems will receive a score based on how well it identifies these different tempi. As the score for each tempo annotation is based on the proportion of tappers that prefer it, the P-Score corresponds to the proportion of perceptual tempi that was correctly identified.

III. TEMPO MODEL

A. Overview

A flowchart of the system is shown in Fig. 2. The system consists of five levels and at each level (described in Secs. III B–III F), data representations are transformed into higher level abstractions. (1) At the Audio Level, harmonic and percussive content of the original waveform is separated

so that the subsequent analysis can be applied to different parts of the audio (Sec. III B). (2) At the Onset Level, the different parts are transformed into the time-frequency domain, and onsets are detected from the computed SF (Sec. III C). (3) At the Interonset level, the onsets are analyzed jointly and different features such as onset density and the strength of the different clusters are computed. At this level IOI histograms are also generated for subsequent analysis (Sec. III D). (4) At the Rhythm Level, the periodicity of the ME is modeled by the cepstroid vector reflecting the periodicity of the IOI histograms. The final tempo is constrained to be at some given multiple of the highest peak in the cepstroid vector. Pulse strength at plausible beat length candidates are calculated by using IOI histograms, weighted based on onset characteristics. Furthermore, the speed of the music is computed based on features such as onset densities and spectral fluctuations (Speed Regression). This estimate is used to find the correct tempo octave (Sec. III E). (5) Finally at the Tempo Level, the tempo is calculated by combining information of period, pulse, and speed in a logistic regression framework (Sec. III F).

The main advantage with a structure consisting of many layers is that each layer can provide invariance to local changes. As an example, the abstraction of an onset is

![Flowchart of the processes used to compute the tempo of an ME. At each level, transformations of the input are used to find a higher level abstraction of the data.](image)

![Two common errors which are perceptually incorrect but would be reported as merely octave errors by the Acc2 measure (dashed arrows). Note also that each example has two tempi (solid arrows), which both could be perceived as correct, depending on the mood and background of the listener, the relationship between sound sources in the mix, the frequency response during playback, etc.](image)
invariant with regard to the exact nature of the spectral change that was produced at the note start. Similarly, in the aggregation of IOI histograms, local variance in IOIs can be removed. Therefore, the subsequent layers can be used to find complex relationships while disregarding perceptually irrelevant factors in the data that have been removed at previous layers.

B. Audio level

At the first step, the original audio file is transformed to separate harmonic and percussive content, based on the method proposed by FitzGerald (2010). In summary, in the original method, harmonic and percussive sources are separated with median filtering in the spectral domain. By applying a median filter across each frame in the frequency direction, harmonic sounds can be detected as outliers. By applying a median filter across each frequency bin in the time direction percussive sounds can be detected as outliers. The median filtered spectrograms (H and P) are used to create soft masks through Wiener filtering. The onset detection function was implemented as described by Elowsson and Fridberg (2013). For the second iteration, the constant-Q transform (CQT), which produces a spectrum with logarithmically spaced frequency bins, was used to filter the percussive waveform from the first iteration. The implementation by Schörkhuber and Klapuri (2010) was applied to produce the CQT time-frequency representation. The frequency resolution was set to 60 bins per octave and each frame was in this case median filtered across the frequency direction with a window size of 40 bins (see also Elowsson and Fridberg, 2013b). Once again the spectrogram was inverted back to the time domain after filtering, resulting in a percussive waveform without harmonic traces.

C. Onset level

The onset detection function was implemented as described by Elowsson and Fridberg (2013). Two independent onset detection functions were used. Percussive onsets are detected from the percussive waveform using an STFT with a window size of 23 ms, and harmonic onsets are detected from the original waveform with a CQT using 60 bins/octave. In both cases the SF is computed from the variation in sound level of the spectral representation of the audio. The SF at each point in time n is given by

\[ SF(n) = \sum_{i=1}^{b} HW(L(n, i) - L(n - s, i)), \]

where b is the number of frequency bins/bands and L is the sound level at each time (n) and frequency (i) position. The variable s is the step size, which is a small integer that varies depending on detection function and HW is a half-wave rectifier function, or for the percussive SF

\[ HW(x) = \begin{cases} 
  x & \text{if } x > 0 \\
  0.2x & \text{if } x \leq 0.
\end{cases} \]

The implication of Eq. (3) is that negative spectral fluctuations have a slight influence on the onset detection function. The reason for half-wave rectification is that the SF will detect increases in energy in some frequency bands, regardless of any decrease in energy in other bands. This enables the SF to detect for example a hi-hat that has been hit shortly after a kick drum. But half-wave rectification will make the

\[ H = OSF_s(X, k, l) \]

\[ P = OSF_s(X, k, l) \]

<table>
<thead>
<tr>
<th>i</th>
<th>k</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–50</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>51–150</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>151–2048</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>1–200</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>201–400</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>401–2048</td>
<td>7</td>
<td>17</td>
</tr>
</tbody>
</table>

TABLE I. The frequency-dependent parameters of the order statistics filters, where k is the order, i is the length and l is the frequency bins. The filtering is used to filter out harmonic and percussive content in the spectrogram.
SF more sensitive to sporadic increases in energy from noise in just a few bands. Therefore, if we also give decreases in energy some importance, the resulting SF curve should become more reliable. The value of 0.2 in Eq. (3) was determined to be a good compromise by observing the effect on relevant audio examples.

For the detection of onsets from instruments with harmonics we applied vibrato suppression, using the technique of max filtering described by Elowsson and Friberg (2013a). In summary, shifts of a peak by 20 cents (one bin), without an increase in sound level are suppressed by subtracting the sound level of each bin of the new frame, with the maximum sound level of the adjacent bins in the old frame. Vibrato suppression for onset detection was proposed independently by several authors at the same time (Lartillot et al., 2013; Elowsson and Friberg, 2013a; Böck and Widmer, 2013).

Finally onsets were retrieved by peak picking on the low-pass filtered SF curve. If two peaks were closer than 70 ms from each other, the peak with the lowest SF-value was removed.

D. Interonset level

To find higher level abstractions of the music a set of representations (features) are computed from the onsets and their relationship to each other. One of these features is the onset density, which is computed for both percussive and harmonic onsets as the number of onsets/s.

By comparing each onset with surrounding onsets within the time-span of 1.5 s, the perceptual impact \( \rho(x) \) of each onset \( x \) was determined (Elowsson and Friberg, 2013a). The onset was given a higher impact if there were no significantly stronger onsets within this time span. Features were once again computed as onset densities but this time weighted according to perceptual impact.

Another set of features are computed from the IOIs within each percussive component. To achieve this, the percussive onsets are clustered into different components with a \( k \)-means clustering based on spectral content, as described by Elowsson and Friberg (2013a). For a comprehensive review of drum clustering techniques and some of its applications, see the summary by FitzGerald and Paulus (2007). As shown in Fig. 3, each onset will be assigned to one of the clusters. The most common IOIs within each cluster are then detected by peak-picking from a smoothed IOI histogram of each cluster.

Two important IOI histograms, the periodicity vector and the beat vector, are also computed at this stage. For these histograms, the IOIs between all pair of onsets that are close enough to each other to represent any perceptually relevant beat or period lengths are used. This is similar to what has been proposed earlier by, e.g., Seppänen (2001). Each bin in the histograms will reflect the population of onset pairs at a particular absolute distance from each other, with a resolution of 1 ms. In addition to the perceptual impact \( \rho \) (outlined above), the spectral similarity of the two onsets \( x \) and \( y \) is also used to compute the contribution of each onset pair in the histograms. Let \( L(x) \) represent the maximum sound level in each band \( i \). Then the spectral similarity of each onset pair \( xy \) is calculated from a band-wise log-spectral distance measure \( D_{\text{LSB}} \), with eight bands spaced approximately an octave apart

\[
D(xy)_{\text{LSB}} = \sum_{i=1}^{8} |L_x(i) - L_y(i)|a(i). \tag{4}
\]

With the weighting parameter array \( a \) the contribution to the distance measure from the different bands is determined, which provides a way of controlling the impact of to the perceptual strength of different frequency bands. The band-wise distance measure has some benefits; specifically, it is less susceptible to variations in pitch in the predominant melody than a bin-wise measure, as long as at least one partial is present in each band. As a final step the histograms (periodicity vector and beat vector) are smoothed with a Hanning window.

For the periodicity vector, the contribution of each onset-pair increases with increasing spectral similarity as well as increasing perceptual impact. Thus, the contribution from each onset pair \( xy \) at the position \( |x - y| \) in the periodicity vector, can be described in principle as

\[
D(xy)_{\text{LSB}} \times (\rho(x) + \rho(y)), \tag{5}
\]

utilizing the perceptual impact described above. An example of the periodicity vector is shown in Fig. 4. Periodicity vectors are used in a large number of tempo estimation algorithms (Gkiokas et al., 2012; Seyerlehner et al., 2007), but it is not common to measure spectral similarity between onsets to compute it.

For the beat vector the contribution of each onset-pair increases with increasing dissimilarity in spectrum as well as increasing onset strength. Thus, the contribution from each onset pair \( xy \) at the position \( |x - y| \) in the beat vector, can be described in principle as

\[
\frac{\rho(x) + \rho(y)}{D(xy)_{\text{LSB}}}. \tag{6}
\]

The idea is based on the observation that the beat length in popular music often corresponds to the distance between the

FIG. 3. (Color online) An example of the clustering of percussive onsets. Three different clusters are found in this case, corresponding to three different combinations of the drums.
kick drum and the snare drum, and that these drums have different spectra. A similar idea has been developed by Peeters and Flocon-Cholet (2012), and they call it spectral balance variation.

It seems reasonable to assume that for the MEs with percussion, repeating patterns in the timbre of the onsets are best captured from the spectrum of the percussive waveform. For the MEs without percussion on the other hand, the spectrum of the original waveform seems better as the percussive waveform will only contain soft traces of the music. Therefore, the MEs are classified as either percussive or harmonic based on the difference in sound level between the waveforms, and the function in Eq. (4) is applied to either the percussive waveform (for percussive MEs) or the original waveform (for harmonic MEs).

E. Rhythm level

We propose a new approach to address the multi-dimensionality of the tempo estimation problem. The assumption behind the approach is that the listener uses speed related features such as onset densities in conjunction with pulse strengths at different metrical positions—that can be given by the most salient periodicity—to infer the perceptual tempo. The outline for the two machine learning steps (Rhythm and Tempo level in Fig. 2) is as follows. (1) Detect the most salient periodicity and use that information to establish a set of plausible tempo candidates. (2) Measure the pulse strength for each plausible tempo candidate. (3) Model the speed of the music as a continuous function. From the representations computed at these steps the final tempo (Tempo Level) is computed: (4) Use the ratio between features of a correct tempo candidate and an incorrect tempo candidate to train a classifier, and utilize the compliance between the proposed tempo candidates and the modeled speed as an additional feature, so that the information from (1), (2), and (3) is propagated to (4).

1. Periodicity

The periodicities in the ME will typically recur periodically as shown in Fig. 4. Unfortunately this makes the periodicity vector a fairly poor representation of the metrical structure of the music. As an example, listeners tapping to the ME in Fig. 4 have been reported to tap periodically at either the triple meter level with a beat length of 0.293 s (first peak in Fig. 4) or with an interval of 0.880 s (third peak in Fig. 4). No listener taps at intervals of 0.587 s (second peak in Fig. 4), although IOIs of that length are common, because it is not a perceptually relevant level of repetition in the music. In conclusion, the periodicity vector is not a correct representation of the perceived metrical structure, as it does not sparsely convey that structure. To alleviate this we introduce the cepstroid vector, computed from the discrete cosine transform (DCT) of the periodicity vector. Our assumption is that if we detect periodically recurring events in the periodicity vector we should be able to create a sparse representation of that vector. Before the DCT is applied to the periodicity vector, it is processed to remove lower magnitudes, which corresponds to less accentuated IOIs or noise. This is accomplished by applying Eq. (7) followed by Eq. (8) to all the elements i of the periodicity vector (V)

\[ V'_i = \max(V_i - 0.15, 0), \]  
\[ V''_i = \max(V'_i - \bar{V'}, 0). \]  

The parts of the vector that are relevant are thus determined by the magnitude in relation to the maximum magnitude and the magnitude in relation to the average magnitude. Subsequently the vector is windowed with the square root of a Hamming window and normalized so that the maximum magnitude is 1. By zero padding the vector (at the end) before finally applying the DCT we get peak positions with a higher resolution.

The spectrum of the DCT represents periodically recurring peaks of the periodicity vector as frequencies. The magnitude of the frequencies corresponds to their salience in the metrical structure. To transform the spectrum back to the time-domain, frequency peaks in the DCT spectrum are detected and their magnitude inserted at the position in the cepstroid vector that corresponds to the fundamental frequency. As a postprocessing step, the cepstroid vector is convolved with the periodicity vector, windowing is applied, and the resulting vector is finally smoothed by convolution with a Hanning window. The motivation behind the post processing is to suppress noise and to favor cepstroids of lengths close to common tempi, as those cepstroid lengths will minimize the amount of \( \frac{4}{3} \) and \( \frac{5}{4} \) octave errors. The cepstroid vector, obtained from the periodicity vector in Fig. 4 through the operations described above, is shown in Fig. 5.

The highest peak in the cepstroid vector will be referred to as the cepstroid (see Fig. 5) and it is important in the proposed system. It is used to restrict which peaks in the beat vector that can correspond to the correct tempo. The assumption behind this model is that the listener understands music by first localizing the most prominent periodicity. After the
most prominent periodicity has been identified the listener proceeds to define the tempo.

A measure with similarities to the cepstroid vector is the autocorrelogram (Lartillot et al., 2013) where the autocorrelation of a periodicity function is calculated. This will also detect periodicities of the periodicity vector. It will however suffer from the same problem as the periodicity vector it will however suffer from the same problem as the periodicity vector with periods recurring at each octave. Also related, Cemgil et al. (2006) has used the DFT of a “tempogram,” Klapuri et al. (2006) has used the DFT of a comb filter with a comb filter to generate a “tempogram,” and Peeters and Peeters and Marchand (2013) used the DFT of the autocorrelation of the SF to try to determine if the tempo was in the region of 110–170 BPM.

2. Pulse

As the next step, the pulse strength in the periodicity vector, the beat vector and the cepstroid vector is computed for different tempo candidates. To generate a set of candidates, the cepstroid from the periodicity detection step is used, and the beat length of the candidates is given by

\[
\text{Cepstroid} \times \left( \frac{1}{2} \right)^n, \quad \text{Cepstroid} \times \left( \frac{1}{2} \right)^n \times \left( \frac{1}{3} \right)^n, \quad n = \ldots, -1, 0, 1, 2, \ldots
\]

Equation (9) implies an assumption that the beat length will have a duple or a triple (times duple) relationship to the most salient periodicity (the cepstroid). This turned out to be true in virtually all MEs in this study, but it will produce erroneous beat length estimations if the cepstroid is detected at the bar level of an odd meter such as \( \frac{5}{4} \) or \( \frac{7}{5} \). Section V offers a discussion about possible extensions to the system. The pulse strengths are defined as the magnitude of the curve in the three different vectors, at the position corresponding to the beat length of each tempo candidate.

3. Speed

The perceptual speed of the music is modeled in a multiple linear regression. As there were no speed annotations in the training data, ground truth speed was approximated from the tempo annotations (see Sec. II B) by

\[
\text{Speed} = \log(\text{AT}_1) + \log(\text{AT}_2)(1 - w).
\] (10)

Note that this is not the same measure as used by Madison and Paulin (2010), Elowsson and Friberg (2013a), and Elowsson et al. (2013), where listener ratings of speed were used. Instead it has a stronger connection to studies on tempo estimation (Gkiokas et al., 2012; Levy, 2011; Chen et al., 2009; Xiao et al., 2008) where speed defines the correct tempo octave. The transformation to the log-domain prior to the regression can be motivated from a perceptual point of view. We perceive a contrast between two tempi based on the ratio between the two, and not based on the absolute difference in BPM (Cemgil et al., 2000). Speed is just an approximation of tempo, positioned at a point somewhere in between the two BPM values of the tempo annotations. Therefore, speed should also be treated logarithmically to properly model tempo. This approach will suppress octave errors equally across the whole tempo range (see Figs. 6 and 7 below). For clarity, speed will be transformed back from the log domain in figures and text to make the connection to tempo clearer. Speed will thus be denoted in BPM in the following description.

The representation of the data that was obtained at the Audio Level, Onset Level, and Interonset Level is used to generate features for the speed regression. They can be divided into four different groups and are presented below:

- The **HP differences** features are based on differences in sound level between the harmonic and percussive part (see Sec. III B). By also computing the maximum sound level over shorter periods, and taking the difference between these values for the waveforms, some complimentary features were extracted.
- **Cluster IOI** features are based on IOIs within clustered components as described in Sec. III D.
- Furthermore, **Onset densities** described in Sec. III D, from the harmonic and percussive onset detection features are included as a group.
- The **Metrical features** are the cepstroid (Sec. III E) and measurements of periodicity patterns in the periodicity vector, from the beginning of the vector to the position of the cepstroid. These patterns can generally be regarded as tendencies for different meters. They are computed by taking the Hadamard product of different metrical templates (scaled to the length of the cepstroid), and the periodicity vector.

Table II shows the results of the speed regression for percussive MEs (the MEs where the system used the...
percussive onsets for the IOI histograms). Features were chosen using forward regression, iteratively adding the best explaining feature. The adjusted $R^2$ is presented for the features of each group as well as for all features in combination. The squared semipartial regression coefficient ($sr^2$) estimates the independent contribution of each group, as it represents the drop in overall performance when the features of a group were omitted and the regression recalculated with the remaining features.

For percussive music, onset densities seem to be the most important. Features based on IOIs of clustered components were also explaining the variance relatively well, but as indicated by the low $sr^2$, the other groups can cover up most of the contribution from that group. In Fig. 6, the modeled speed is compared to the ground truth speed for percussive MEs.

The regression for harmonic MEs (the MEs where the system used the harmonic onsets for the IOI histograms) is presented in Table III. As in Table II, the features have been divided into different groups. Onsets were not clustered into different components for the harmonic MEs, so that group is omitted.

Onset densities were once again the most important group. There was just a single periodicity feature, a measure of triple meter probability. In Fig. 7, the modeled speed is compared to the ground truth speed for harmonic MEs.

The MEs that were classified as harmonic are noticeably slower than the percussive MEs. The mean annotated speed (as converted back from the log domain) is 81 BPM for the harmonic MEs and 110 BPM for the percussive MEs.

**F. Tempo level**

In the final step, the plausible tempo candidates (see Sec. III E 2) are evaluated using a logistic regression framework. Logistic regression is a probabilistic classification model, where a logistic (sigmoid) function is used to map the range of negative infinity to positive infinity into the range of 0 to 1. For each tempo candidate ($C$), the representation of the data that was obtained at the Rhythm Level is used to generate features for the logistic regression classification. The features are computed from the beat length $C_{len}$ of the candidate and are presented below:

- The Period features are based on the ratio between $C_{len}$ and the cepstroid. As an example, if $C_{len}$ is a third of the cepstroid, the “1/3-feature” is set to 1 and otherwise 0.
- The Pulse features correspond to the magnitude in the different IOI histograms (e.g., beat vector) at $C_{len}$. As an example, if the magnitude of the beat vector at $C_{len}$ is 0.24 in the normalized vector, the “beat vector-feature” is set to 0.24.
- The Speed features are based on the distance between $C_{len}$ and the computed speed of the song, as well as general parameters to adjust the model for faster or slower songs. To compute the distance, a log-normal curve (the Speed window) is used, centered at the position given by the estimated speed (see Fig. 8, below). The window is log-normal in accordance with the notion of speed as a logarithmic perceptual phenomenon. As an example, if the computed speed for the ME is 84 BPM and the tempo which corresponds to $C_{len}$ is close to 84 BPM the “speed-feature” will be high.

Note that several features are computed for each category, with slight modifications, so that the best feature

**TABLE II. The linear regression that models the speed of the percussive MEs. Features have been divided into feature groups.**

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Number of Features</th>
<th>Adj $R^2$</th>
<th>$sr^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP difference</td>
<td>4</td>
<td>0.175</td>
<td>0.063</td>
</tr>
<tr>
<td>Cluster IOI</td>
<td>4</td>
<td>0.314</td>
<td>0.036</td>
</tr>
<tr>
<td>Onset densities</td>
<td>4</td>
<td>0.357</td>
<td>0.087</td>
</tr>
<tr>
<td>Metrical</td>
<td>6</td>
<td>0.179</td>
<td>0.066</td>
</tr>
</tbody>
</table>

FIG. 6. (Color online) Modeled speed for 586 percussive MEs versus ground truth speed (as estimated by the tempo annotations). The dashed line indicates the position where an octave error would be produced in duple meter and the dotted line indicates octave error for triple meter.

FIG. 7. (Color online) Modeled speed for 240 harmonic MEs versus ground truth speed (as estimated by the tempo annotations). The dashed line indicates the position where an octave error would be produced in duple meter and the dotted line indicates octave error for triple meter.
representation can be chosen by the classifier. After the features have been extracted, tempo candidates that are implausible (magnitude < 0.1 in the beat vector) are removed.

In Fig. 8, the feature extraction process is visualized. The cepstroid of the ME is 0.880 s tempo candidates have been generated at beat lengths of 0.293, 0.440, 0.880, and 1.760 s corresponding to the cepstroid multiplied by \( \frac{1}{3}, \frac{1}{2}, 1, 2 \), respectively. The computed features (represented by the vertical position of the circles) from the periodicity vector, the cepstroid vector and the speed are shown. Note that the speed window reaches its maximum at the position given by the estimated speed which happens to correspond to a beat length just below 0.88 s. The features have been developed to approximate a ratio scale, i.e., a two-fold increase should correspond to a two-fold increase in probability.

The logistic regression classifier is not directly trained to classify if a tempo is correct or not but trained instead to create training data, all possible pairs of tempo candidates \( \{C_1, C_2\} \) are examined, and the classifier is trained to classify which of the two candidates in each pair that is most likely to correspond to a correct tempo. Let us denote a tempo candidate that corresponds to an annotated tempo AT\(_1\) or AT\(_2\) (see Sec. II B) as \( C_1 \), and a tempo candidate that does not correspond to an annotated tempo as \( C_0 \). If none of the two tempo candidates in a pair is correct \( \{C_0, C_0\} \), the pair is dismissed. If one candidate is correct and one is incorrect \( \{C_1, C_0\} \), and if the annotation that was matched to \( C_1 \) has a weight \( w > 0.2 \) (see Sec. II B), training data is generated for the logistic regression. Denote the computed features of each tempo candidate by \( C_{\text{feat}} \). Each pair of tempo candidates generates two dependent variables \( Y \) (0 and 1), and two independent variable sets \( X \), by computing what can be regarded as a log-likelihood of each feature \( i \),

\[
Y = 1, \quad X_i = \log \left( \frac{C_{\text{feat}(i)}}{C_{\text{feat}(i)}} \right), \tag{11}
\]

\[
Y = 0, \quad X_i = \log \left( \frac{C_{\text{feat}(0)}}{C_{\text{feat}(1)}} \right). \tag{12}
\]

By letting the logistic regression classifier model a binomial distribution, the classifier is also trained on the pairs where both tempo candidates are correct \( \{C_1, C_1\} \). For these pairs the dependent variables \( Y \) are instead associated with the binomial distribution defined by \( w \), the weight of the more salient tempo. We let \( Y \) be \( w \) and \( 1-w \), respectively, which should be appropriate for the data as \( w \) corresponds to the percentage of people tapping to AT\(_1\) in relation to AT\(_2\). Let the most salient candidate be defined as \( C_w \) and the less salient candidate be defined as \( C^{1-w} \), then the log-likelihood of each feature \( i \) was computed by

\[
Y = w, \quad X_i = \log \left( \frac{C_{\text{feat}(i)}}{C^{1-w}_{\text{feat}(i)}} \right), \tag{13}
\]

\[
Y = 1-w, \quad X_i = \log \left( \frac{C^{1-w}_{\text{feat}(i)}}{C_{\text{feat}(i)}} \right). \tag{14}
\]

Finally, all the collected dependent variables \( Y \) and the independent variable sets \( X \) are used to train the logistic regression classifier.

At run-time, AT\(_1\) and AT\(_2\) are unknown (when testing the system on the test set or running the system on new MEs). At this point, all pairs of plausible candidates are accumulated, and the variable sets \( X \) are generated in a similar way as described above. In the final classification, the derived weights from the training of the logistic classifier are applied between 0 and 1. The candidate with the highest mean score (from all pair-wise evaluations) is proposed as T\(_1\) and the candidate with the second highest score is proposed as T\(_2\). For the ME in Fig. 8, the cepstroid-candidate received a score of 1.00, the \( \frac{1}{3} \)-candidate received a score of 0.31 and the \( \frac{1}{2} \)-candidate received a score of 0.2. The candidate at \( \frac{1}{2} \) the cepstroid was removed according to the conditions above (beat vector magnitude <0.1), and therefore never evaluated for training or testing. This exclusion of unlikely candidates from classification has the effect that the weights will be optimized for classifying the more difficult pairs.

Why did we choose to train the classifier by comparing pairs of tempo candidates? The idea behind this approach is that the probability that a certain tempo candidate is correct also depends on the features of other candidates for the same song. For example, some songs will exhibit clear
periodicities at several metrical levels whereas other songs will not exhibit strong periodicities at any level. Thus, it is not the level of periodicity that defines if a tempo candidate is correct or not, but the level of periodicity in relation to the periodicity of other plausible tempo candidates within the same song. It would have been possible to account for this in other ways as well. One strategy could have been to calculate global features for each ME (e.g., the mean periodicity of all tempo candidates) so that they can be added to the regression for implicit normalization. The ratio between features was chosen because it gives fewer features, and features that are easy to interpret. Note that some normalization is still taking place (i.e., each histogram/vector has a max value of 1).

1. Contribution from tempo features

In Tables IV and V the contribution from the features used for the logistic regression is presented. Features were chosen with a forward regression, iteratively adding the best explaining feature in each step. Once again the features have been grouped to provide better overview. Let the deviance $D$ be the generalization of the residual sum of squares in logistic regression. The proportional reduction in deviance from the logistic regression is given by

$$R^2 = \frac{D_{null} - D_{model}}{D_{null}}, \tag{15}$$

where $D_{model}$ is the deviance of the model and $D_{null}$ is the deviance when there are no predictors. $R^2_L$ is a pseudo-$R^2$ measurement and it is analogous to the $R^2$ measurement in linear regression in that it represents the amount of explained variance (Cohen et al., 2003, pp. 500–502). The ratio of pairs that were correctly classified, defined as $\{Y-score < 0.5\}$, is given by $R^2_L$ in Tables IV and V. In Table IV the contribution from features for percussive MEs are presented. As in Tables II and III the semipartial regression coefficient $sr^2_L$ estimates the independent contribution of each group.

Evidently the speed features are the most important to classify tempo. The strength at different pulse levels is also contributing, whereas the addition of the Period group to the other two groups only provides an additional 1.5% of explained variance. In Table V, the contribution from features for harmonic MEs are presented.

The speed features are once again the most important. The pulse strength is however also a relevant group; if it is omitted from the regression about 11% of explained variance is lost. The contribution from the period group is once again the smallest, with a 4% reduction in explained variance when omitted. Two features were always chosen first and second when training the classifier with features from all groups. The first chosen feature was computed from the Speed window and the second feature was the magnitude in the beat vector.

IV. RESULTS

A. Evaluation method and datasets

To evaluate the system a tenfold cross-validation was used. The whole system (i.e., both the linear regression used to compute speed and the logistic regression used to find the correct tempo) was trained with 90% of the MEs and the remaining 10% were evaluated. By also repeating the cross-validation procedure ten times and taking the mean, consistency was ensured.

We will report results for two publicly available datasets with annotated tempo, Ballroom and Songs, which were used in the ISMIR tempo induction contest of 2004 (Gouyon et al., 2006). The tempo annotations and audio files of the Ballroom dataset originated from a homepage providing online ballroom dancing lessons (Ballroomdancers, 2006). These annotations were later double-checked by Simon Dixon and made available to other researchers (Gouyon et al., 2006). The 465 MEs in the Songs dataset contains a wide variety of genres, including Balkan and Greek (144 MEs), Classical (70 MEs), Rock (68 MEs), and Electronica (59 MEs). The tempo annotations of this dataset were computed from the median distance between beat markers that had previously been annotated by a professional musician (Gouyon et al., 2006). We note that Acc1 results on the datasets have been reported for many years by a large number of tempo estimation algorithms. Therefore, Acc1 results (with original annotations) serve as a good point of reference and we will report results for the system by this standard. It should however be noticed that the Ballroom dataset is extremely unbalanced, containing only a few genres with the MEs of each genre being largely identical in instrumentation and tempo. This could give template-based methods an advantage if they have been trained on the set in a cross-validation setup. The annotations are however accurate, in so far that the notated tempo is accurately provided (see Sec. I for a distinction between notated and perceptual tempo). The Songs dataset is more balanced but with less accurate annotations.

We will also report results for the test set of the MIREX Audio Tempo Estimation task, which is not available to submitters. The set consists of 140 songs of varying genres,

<table>
<thead>
<tr>
<th>Harmonic MEs</th>
<th>Number of features</th>
<th>$R^2_L$</th>
<th>$sr^2_L$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>3</td>
<td>0.672</td>
<td>0.378</td>
<td>0.931</td>
</tr>
<tr>
<td>Pulse</td>
<td>2</td>
<td>0.355</td>
<td>0.112</td>
<td>0.818</td>
</tr>
<tr>
<td>Period</td>
<td>7</td>
<td>0.303</td>
<td>0.038</td>
<td>0.791</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percussive MEs</th>
<th>Number of cases 2644</th>
<th>$R^2 = 0.895$</th>
<th>$sr^2_L$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>4</td>
<td>0.791</td>
<td>0.493</td>
<td>0.963</td>
</tr>
<tr>
<td>Pulse</td>
<td>4</td>
<td>0.350</td>
<td>0.058</td>
<td>0.814</td>
</tr>
<tr>
<td>Period</td>
<td>5</td>
<td>0.163</td>
<td>0.015</td>
<td>0.750</td>
</tr>
</tbody>
</table>

TABLE V. The logistic regression that is used to identify the correct tempo for the harmonic MEs. Features have been divided into feature groups.
tempo and meter (Moelants and McKinney, 2004; McKinney and Moelants, 2004) annotated and evaluated as described in Sec. II B. The competition offers a formal performance benchmark for tempo estimation software (Levy, 2011). It is especially useful that the test set is hidden, which prevents any overfitting on the MEs.

Unfortunately, there are no datasets annotated with both AT1 and AT2 available for training, except the 20 MEs released for submitters to the competition. Therefore, ground truth annotations with two tempi and a corresponding weight, in accordance with the MIREX Audio Tempo Estimation task, have been developed for this study. Over 800 MEs were selected for the dataset, most taken from the Ballroom and Songs dataset, with 50 MEs from a study by Madison and Paulin (2010) and the already annotated 20 MEs from the MIREX training set. As the existing annotations for the Songs and Ballroom datasets do not correspond to perceptual tempo they could not be used. Instead, the MEs were annotated by the first author and care was taken to imitate the perceptual assessments in the annotations and weights of the MIREX training set. A few annotations were adjusted after the initial assessment. The training set is called the Combined dataset and will be available for research purposes.1 We will report results for this training set according to both the P-Score measurement and the Acc1 estimation.

### B. Overall results

The results on the four datasets are shown in Table VI. The MIREX results are from the evaluation in the 2013 MIREX task. As the Ballroom and Songs datasets do not contain annotations for P-Score it cannot be reported for them, and unfortunately Acc1 is not reported from the MIREX competition.

### C. Results for the combined dataset

In Table VII, the detailed results for the Combined dataset are presented. Results are broken down into percussive and harmonic MEs, based on the classification by the system. Recently a number of studies have examined songs with a well-defined tempo (Peeters and Flocon-Cholet, 2012; Levy, 2011). For the Combined dataset the weight w can be used as an estimate of the confidence of the most prominent tempo. The tempo of an ME where one of the annotations has a weight above 0.9 is likely unequivocal. Statistics for these MEs are presented in the category Acc1w. The different types of errors are listed, and they have been divided into three categories. Err represents the erroneous estimation for an ME which has no relevant relationship to the correct tempo. These errors are generally due to incorrect cepstrum estimations. OctA is the cases where the incorrect tempo octave was chosen by the system. These results are broken down for the four octave error types (1/2, 1, 2, 3). For P-Score, octave errors are typically occurring when one of the two estimated tempi is correct and one is incorrect. For these cases the octave relationships are based on the ratio between the tempo that is incorrect and the tempo that is correct (e.g., 2, means that the incorrect estimation was twice that of a correct estimation). OctB represents errors that are smaller than an octave but with a relevant periodic relationship (an estimation which is 1/3 or 2/3 of the correct tempo).

The tempo of the harmonic MEs was generally harder to estimate than the tempo of the percussive MEs and this phenomenon will be discussed in Sec. V. The reported P-Score for the Combined dataset was 0.963. The errors are evenly distributed over the three error categories. The Acc1 measure is correctly estimated for about 89% of the MEs. The main source of errors is octave errors, with about 8% of the errors stemming from the OctA type and 2% stemming from the OctB type. When only the MEs with a well-defined tempo were evaluated, about 93% of the MEs were correctly estimated. The improvement is explained by a drop in the number of octave errors of type OctA.

How does the confidence of the listeners’ tempo annotation relate to the Acc1 estimation? When w for an ME is 0.51, both tempi (AT1 and AT2) have been chosen almost an equal number of times when averaged over listeners. Naturally, when the two tempi are almost equally valid, it is difficult for a tempo estimation algorithm to decide which of the two tempi that is the most prominent. It should be possible to use w as a measure of the confidence with which the most prominent tempo can be identified in an ME. By taking the average w of all MEs in the dataset we get an estimate of the expected human performance level in estimating Acc1 for them (wavg). As an example, a dataset with an average w of 0.75 implies that the listeners on average are able to find the most prominent tempo 75% of the time. Therefore, when the average w is 0.75, it should be hard to reach an Acc1 above 75% for any tempo estimation algorithm that is run on the dataset. Altering the composition of the dataset by iteratively removing the MEs with a w below wlim, enables us study the relationship between the listeners’ performance estimation.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of MEs</th>
<th>Results</th>
<th>OctA</th>
<th>1/2</th>
<th>2</th>
<th>3</th>
<th>OctB</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-Score</td>
<td>586</td>
<td>0.981</td>
<td>0.007</td>
<td>0.011</td>
<td>0.002</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>Harm</td>
<td>240</td>
<td>0.918</td>
<td>0.025</td>
<td>0.018</td>
<td>0.008</td>
<td>0.003</td>
<td>0.007</td>
</tr>
<tr>
<td>Total</td>
<td>826</td>
<td>0.963</td>
<td>0.012</td>
<td>0.013</td>
<td>0.004</td>
<td>0.007</td>
<td>0.002</td>
</tr>
<tr>
<td>Acc1</td>
<td>586</td>
<td>92.0</td>
<td>0.2</td>
<td>7.3</td>
<td>4.9</td>
<td>2.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Harm</td>
<td>240</td>
<td>81.2</td>
<td>2.5</td>
<td>11.0</td>
<td>1.3</td>
<td>4.0</td>
<td>5.6</td>
</tr>
<tr>
<td>Total</td>
<td>826</td>
<td>88.9</td>
<td>0.8</td>
<td>8.4</td>
<td>0.4</td>
<td>4.6</td>
<td>3.2</td>
</tr>
<tr>
<td>ACC1w</td>
<td>371</td>
<td>94.8</td>
<td>4.4</td>
<td>2.6</td>
<td>1.8</td>
<td>2.9</td>
<td>5.2</td>
</tr>
<tr>
<td>Harm</td>
<td>127</td>
<td>88.6</td>
<td>1.6</td>
<td>4.6</td>
<td>1.8</td>
<td>2.9</td>
<td>5.2</td>
</tr>
<tr>
<td>Total</td>
<td>498</td>
<td>93.3</td>
<td>0.4</td>
<td>4.5</td>
<td>2.4</td>
<td>2.1</td>
<td>1.9</td>
</tr>
</tbody>
</table>

1 We will report results for this training set according to both the P-Score measurement and the Acc1 measurement.

---

1. A. Elowsson and A. Friberg: Modeling the perception of tempo


3173
level \( w_{avg} \) and the performance of the estimation algorithm \( Acc_{1w} \). This is shown in Fig. 9.

The performance of the system seems to be on par with the human accuracy level \( w_{avg} \). The drop in performance (relative to the human performance level) for the MEs with the most well defined tempo may be due to the increased negative impact of any periodicity errors on the relative performance, or due to the smaller number of training points available. As \( w \) approaches 1, the number of MEs that can be used for training is reduced. As a result, the few pairs that are left are perfectly classified in the training set but the logistic regression still becomes unstable for the test set. Therefore, only results for certainties up to 0.95 are presented. In this article, \( w \) of the Combined dataset was annotated by only one listener so it is a fairly crude estimate, but it should nevertheless be a fairly relevant approximation of human level performance.

D. Comparison with previous results

1. Ballroom and Songs

Figure 10 shows the performance of the system in comparison with the 20 best systems described in the literature for the Ballroom dataset. The proposed system achieves a classification accuracy \( Acc_1 \) of 93.0\%, which is the best so far in comparison with all other methods described in the literature for the dataset.

The MEs that the system treated as harmonic had a classification accuracy of 88.9\% and the percussive MEs had a classification accuracy of 95.4\%. It should be mentioned that although cross-validation was used, any machine learning approach trained on a subsample of a dataset will be somewhat tuned to the characteristics of the dataset. The systems denoted by (+) in Fig. 10 were presented by Gouyon et al. (2006), and they were evaluated in the MIREX tempo estimation task of 2004. Comparisons should therefore be done with caution, as they did not have access to the dataset beforehand. Furthermore, the template based methods which do not use rescaling\(^2\) (*) may, as noted by Xiao et al. (2008), be extra prone to overfitting an unbalanced dataset such as Ballroom.

The statistical significance of the difference between the current model and the second best (Gulati and Rao, 2010) in

![FIG. 9. (Color online) The performance of the system (\( Acc_1 \)) in comparison with the annotated human performance level (\( w_{avg} \)).](image)

![FIG. 10. (Color online) The result for the proposed system (shaded bar at the top) on the Ballroom dataset in comparison with the 20 best systems proposed in the literature. For comparison (*) denotes the template based (without rescaling) and genre based methods, and (+) denotes the methods from the MIREX tempo estimation task of 2004 (Gouyon et al., 2006). Systems are denoted by first author of the corresponding literature reference, except for the systems that were presented in the meta-evaluation of the MIREX tempo estimation task by Gouyon et al. (2006). When several systems were proposed by the same author in the same study they have been differentiated with a superscripted number.](image)

![FIG. 11. (Color online) The result for the proposed system (shaded bar at the top) on the Songs dataset in comparison with the 20 best systems proposed in the literature. For comparison (*) denotes the template based (without rescaling), (+) denotes the methods from the MIREX tempo estimation task of 2004 (Gouyon et al., 2006) and (C) denotes the algorithms tested by Zapata and Gómez (2011). Systems are denoted by first author of the corresponding literature reference, except for the systems that were presented in the meta-evaluations by Gouyon et al. (2006) and Zapata and Gómez (2011). When several systems were proposed by the same author in the same study they have been differentiated with a superscripted number.](image)
Fig. 10 was tested with Pearson’s chi-square test, as the track-by-track results necessary for McNemar’s test were not available. The chi-square test was significant at the p < 0.05 level (p = 0.0025).

Figure 11 shows the performance of the system in comparison with the 20 best systems described in the literature for the Songs dataset. The proposed system achieves a classification accuracy of 77.3%, which is better than any other method described in the literature for this dataset.

We note that the difference between the proposed systems and other systems in the literature seems to be bigger for the Songs dataset. The difference is likely due to the more balanced distribution of tempi in the dataset, making template-based methods without rescaling less successful. The chi-square test of the relationship between results of the proposed system and a combination of methods by Zapata and Gómez (2011) produced p = 0.00007 thus statistically significant at the p < 0.05 level.

E. MIREX

The system participated in the MIREX tempo estimation task of 2013 (MIREX, 2013) The reported P-Score of the system was 0.857. As shown in Fig. 12 below, this is the highest results reported so far in the competition (2006–2013), among 35 contributions.

Significance was tested with bootstrapping for paired samples as the song-by-song difference between algorithms is not normally distributed. Testing the system against the other systems independently, the result was statistically significant at the p < 0.05 level for all but the four top contestants, with p = 0.144 in comparison with the second best system. A bigger test-set could be useful to differentiate the algorithms at a statistically significant level. Two statistics that are reported from the competition is the ratio of estimating at least one tempo correct and the ratio of estimating both tempi correct. The system is the fourth best at finding at least one correct tempo (94% of the MEs), with the best system finding one correct tempo in 96% of the MEs. The system is best at finding both correct tempi (69% of the MEs), with the second best system finding both tempi in 63% of the MEs. Using McNemar’s statistical test, neither of these differences are statistically significant at the p < 0.05 level.

V. CONCLUSIONS, DISCUSSION, AND FUTURE WORK

A tempo estimation system has been proposed which achieves the highest results so far in a formal benchmarking test as well as two widely adopted test sets. The good results indicate that the proposed method accurately models tempo. The continuous speed representation seems to be an effective means of reducing octave errors in tempo estimation. Although the effect of the introduction of a band-wise log-spectral distance measure for onsets cannot be directly quantified, the measure is likely important, as it gives us perceptually relevant IOI histograms at the Interonset level. The cepstroid is a robust way of finding a salient periodicity from which the tempo can be identified. With the perceptually motivated multi-layered structure the system seems to have become fairly invariant to local changes in the training data, as indicated by the high results for the fairly balanced Songs data set.

The system is better at tracking tempo for music with percussive instruments. Several explanations are plausible. It seems reasonable to assume that tonal aspects define the meter to a much higher degree for music without percussive instruments. For some of these MEs, the rhythmic dimension will not provide enough information to correctly estimate the tempo. And as the system does not track pitch or chord progressions, it cannot infer periodicities in the pitch dimension of the melodic structure. Therefore, it is not surprising that the tempo of songs without percussion is harder to estimate. We conclude that efforts in the area of polyphonic transcription may improve the system. Therefore, we intend to pursue that challenge as the next step of the tempo estimation project.

To achieve better tempo estimations the system must also be able to handle odd and changing meter as well as fluctuating tempo. Consequently it is necessary to track music over time, and we see beat tracking as a natural task to pursue. We expect the cepstroid vector to be useful because of its sparse representation of the musical structure, which will be less sensitive to changing tempo. The cepstrogram (tracking the cepstroid vector over time) can be regarded as an alternative to the log-spaced autocorrelation function (Jensen et al., 2009). Furthermore the dependency on a correct cepstroid estimation should be loosened. The concept with the cepstroid representing a perceptual cue to the correct tempo will however remain.

New ground truth annotations have been created for a combination of MEs that are used in the community. The annotation corresponds to perceptual tempo in compliance with the MIREX Audio Tempo Estimation task. The P-Score for this training set was about 0.96 and the Acc1 was about 89%.
The model with continuous speed proposed in this study could be used as an extension to other tempo estimation algorithms, either to evaluate if an existing estimation is in the correct octave, or to add an octave estimate \( T_2 \). In both cases, the estimated speed would be compared to the existing musical accentuation should be measured in a continuous matter instead of using discrete onsets (Eronen and Klapuri, 2010). We believe this to be unfortunate, not only because of the usefulness of the discrete measure in various other MIR-tasks. The onset reduces complexity by transforming the continuous accentuation curve to a higher level representation of the music. The transformation is perceptually motivated; the listener presumably understands significant parts of the soundscape as discrete events from different sources (the pluck of a string, the beat of a snare drum, etc.). Of course, each onset should be thoroughly analyzed (accent strength, pitch height, length, accentuation in relation to surrounding onsets, timbre in relation to surrounding onsets, etc.) so that its role in the music as a whole is fully understood before its contribution to different aspects of rhythm perception can be modeled.

**ACKNOWLEDGMENTS**

This work was supported by the Swedish Research Council, Grant numbers 2009-4285 and 2012-4685.

1With the hope that the annotations can be of use to researchers in the field they are available (at request) for research purposes, by contacting elow@kth.se.

2The study by Schuller et al. (2008) which is included in this group uses the tempo distribution of the nine genres in the Ballroom dataset to classify tempo. For the results of the study, \( Acc_1 \) and \( Acc_2 \) seem to have been reversed.

3No results for template based methods without rescaling have, as far as we know, however been presented for the Songs dataset since the initial study (Seyferlehner et al., 2007).

4In their study, their evaluated methods were combined to produce an “optimal” tempo estimate. This score (65.4) is given by Zapata and Gómez (2011) in the Fig. 8.

5As of the submission date of this article (July 2014). However, before it was published, a submission by Sebastian Böck received a higher P-Score.

6Results are not reported on a song-by-song basis for systems from 2006. They were however not close to reach the statistical significance level.


