MUSIC TEMPO ESTIMATION AND BEAT TRACKING BY APPLYING SOURCE SEPARATION AND METRICAL RELATIONS

Aggelos Gkiokas [1,2], Vassilis Katsouros [1] , George Carayannis [2] and Themos Stafylakis[1]

[1]Institute for Language and Speech Processing / R.C. "Athena"
[2]National Technical University of Athens
{agkiokas, vsk, themosst}@ilsp.gr, carayan@softlab.ece.ntua.gr

ABSTRACT

In this paper, we present tempo estimation and beat tracking algorithms by utilizing percussive/harmonic separation of the audio signal, in order to extract filterbank energies and chroma features from the respective components. Periodicity analysis is carried out by the convolution of feature sequences with a bank of resonators. Target tempo is estimated from the resulting periodicity vector by incorporating metrical relations knowledge.

Tempo estimation is followed by a local tempo refinement method to enhance the beat-tracking algorithm. Beat tracking involves the computation of the beat saliencies derived from the resonators responses and proposes a distance measure between candidate beats locations. A dynamic programming algorithm is adopted to find the optimal "path" of beats. Both tempo estimation and beat tracking methods were submitted on MIREX 2011, while the tempo estimation algorithm was also evaluated on ISMIR 2004 Tempo Induction Evaluation Exchange Dataset.

Index Terms— tempo estimation, beat tracking, chroma features, periodicity analysis

1. INTRODUCTION

Most tempo estimation systems involve a three step process. Firstly, a midlevel representation is extracted from the audio signal that supposes to capture all the rhythmic relevant events within a song excerpt. Spectral Complex Difference [1] and band energies evolution over time [2-4] as well as onset detection functions [5] [6] are widely used for tempo estimation. However, there is evidence that continuous representations perform better [7]. Alonso et al. in [8] proposed a system that estimates the tempo by decomposing the music signal sub-bands into harmonic and noise components. Recent approaches [9] [10] utilize chroma related features as midlevel representations.

Subsequently, the midlevel representation -also referred as accent features- is parsed to extract a periodicity vector, i.e. a vector containing the saliencies of the target periods. This is usually achieved by self-similarity approaches as the Autocorrelation Function (ACF) [1], [11] or by processing accent features by a bank of resonators [3] [4] [10]. The final step involves the selection of the target tempo. This is usually achieved by peak picking in the periodicity vector. In [3], prior knowledge such as prior period distributions and metrical relations are used in a probabilistic model to choose the correct tempo. Recent approaches [9], [11] incorporate classification techniques.

For the beat tracking task, there is a variety of methods reported in the literature. Dixon in [5] utilizes agents to track beat sequences hypotheses in a multiple hypothesis search. Davies and Plumbley [1] utilize a two-state model to handle discontinuities in beats caused by switching metrical levels. In [12] an onset detection system based on bidirectional long short-term memory neural networks was extended to a beat tracking system that performed best in MIREX 2010 Audio Beat Tracking task. Ellis [2] formulates beat tracking as an optimization problem and employs dynamic programming to solve it. Peeters and Papadopoulos [13] propose beat templates to track beats and downbeats through a probabilistic framework.

In this paper, we extend the tempo estimation algorithm presented in [10] by utilizing harmonic/percussive separation of the audio signal, to enhance feature extraction. In addition, we exploit metrical relations knowledge, and the introduction to the notion of "fundamental tempo", in order to estimate the target tempo from the resulting periodicity vector. Tempo estimation is followed by a beat-tracking algorithm, which defines distances between candidate beat locations and adopts a dynamic programming algorithm to the find the optimal "path" of beats.

2. TEMPO ESTIMATION

2.1. Feature Extraction

To extract the desired features, the constant Q transform (CQT) of the audio signal is calculated, using 12 bins per octave, with 25Hz/5kHz minimum/maximum frequencies, and a Hanning window with half overlap. Frequency bins are aligned to the western scale musical pitches and are rescaled by bicubic interpolation/decimation to a 200Hz frame rate, resulting the log-frequency spectrogram

$\mathbf{S} = \{|W_{i,f}|\}$ where $W_{i,f}$ denotes the CQT and *i, f* denote the time and frequency bin indices respectively.

To enhance the extraction of the desired features we apply to **S** the harmonic/percussive separation algorithm proposed in [14]. We used soft masking and γ=0.6. We denote **H** and **P** the harmonic/percussive components derived from **S** respectively.

For each time index *i* we sum up the amplitudes of the frequency bins in **H** that correspond to the 12 semitones of western musical scale in order to compute the 12-dimensional chroma vector $\mathbf{x}_{ch}$:

$$x_{ch}^{k}[i] = \sum_{f_k \in F_k} H_{i,f_k}, \quad k = 1..12 \tag{1}$$

where F_k are the bins corresponding to tone *k*.

In a similar way, filterbank energies are extracted from **P** with 8 logarithmically scaled, equal bandwidth and overlapping triangular filters. We denote filterbank energies as $\mathbf{x}_{fl}$.

2.2. Periodicity Analysis

Feature vectors are differentiated and convolved with a bank of resonators as in [10] in the range of [30,500] bmp. The resonator outputs are segmented using a rectangular window of *Q* times the target tempo period and one period overlap. The value for *Q* was set to 8. We compute the salience of tempo *t* in segment *s* as the maximum amplitude value of the corresponding resonator output in segment *s*.

The above process is used to calculate the "tempogram" matrices $TG^{v}(t,s)$ for each feature vector *v*. The constant Q value property has the effect of fewer segments for smaller tempos, thus the rows of $\mathbf{TG}^{v}$ are time-wrapped to have the same size that is equal to the size of the faster tempo. Then, for each segment index *s,* tempograms are summed for each feature class (filterbank/chroma) independently resulting the matrices $\mathbf{TG}^{ch}$ and $\mathbf{TG}^{fl}$ for chroma and filterbank features respectively. To estimate the global periodicity vector for the whole excerpt denoted by $\mathbf{T}_{gl}$, $\mathbf{TG}^{ch}$ and $\mathbf{TG}^{fl}$ are summed across all segments and then multiplied:

$$T_{gl}(t) = (\sum_{s} TG^{fl}(t,s))(\sum_{s} TG^{ch}(t,s)) \tag{2}$$

2.3. Choosing the Correct Metrical Level

Our method makes the assumption that the peaks in the global periodicity vector that are musically relevant to the ground truth are integer multiples of a certain tempo value. We compute the fundamental tempo T_0 as

$$T_0 = \arg\max_{t} \{\sum_{k=1}^{4} T_{gl}(kt)\} \tag{3}$$

Then we expect that $\mathbf{T}_{gl}$ has peaks at target tempo as well as at integer multiples of T_0. Analogously to the {meter, tactus, tatum} hierarchical model for beat relations, in our approach we consider a model of two tempi $\{T_s, T_f\}$ values (slow, fast) under the assumption that T_s is the more perceptually relevant, while T_f is more likely to be double, triple or quadruple of T_s. We define the joint salience $J_s(T_s,T_f)$ of T_s, T_f as:

$$J_s(T_s,T_f) = [T_{gl}(T_s) + T_{gl}(T_f)] \cdot \sum_{i=2..4} e^{-(T_f/T_s - i)^2/(\rho i)^2} \tag{4}$$

It is clear that J_s becomes greater as the saliencies of T_s, T_f, increase, and when the latter is close to the double, triple or quadruple of the first. The final tempo *T* is the T_s that maximizes J_s and is a multiple of T_0, i.e.:

$$T = \arg\max_{iT_0} \{J_s(iT_0, kT_0), iT_0, kT_0 \in \{30,..500\}\} \tag{5}$$

3. BEAT TRACKING ALGORITHM

3.1. Beat Candidates and Saliencies

As beat candidates we consider the peaks of the resonator responses $\mathbf{r}_{ch,T}^{k}$, $\mathbf{r}_{fl,T}^{m}$ corresponding to the found tempo *T* for all feature sequences $\mathbf{x}_{ch}^{k}, \mathbf{x}_{fl}^{m}$ respectively. Because the phase responses of the resonators cause lags between the peaks of the feature sequences and the peaks of the responses, we modify the resonator of tempo *T* to have zero response phase and magnitude response equal to the square of the magnitude responses of the resonator used at the tempo estimation step. This produces synchronized feature sequences and resonator responses. We denote the time instances of beat candidates as $\{b_j\}, j = 1..N$ and the corresponding saliencies as s_j^b, which are computed as:

$$s_j^b = \hat{\mathbf{r}}_{fl,T}(b_j) + \hat{\mathbf{r}}_{ch,T}(b_j) \tag{6}$$

where

$$\hat{\mathbf{r}}_{c,T}(k) = \sum_{i} \mathbf{r}_{c,T}^{i}(k) / \max_{s} \left(\sum_{i} \mathbf{r}_{c,T}^{i}(s) \right), \; c = \{fl, ch\} \tag{7}$$

3.2. Inter-beat Distances

The distance of two beat candidates must increase as the time interval between them diverges from the target period τ_T and must decrease as the candidate next beat salience increases. We propose the distance between b_i, b_j with $b_j > b_i$, as follows

$$d(b_i,b_j) = \gamma \cdot d_T(b_i,b_j) - (1-\gamma) s_j^b \tag{8}$$

where

$$d_T(b_i, b_j) = 1 - \exp\{-\frac{1}{\sigma^2}\ln^2\left((b_j - b_i)/\tau_T\right)\} \quad (9)$$

Parameter $\gamma \in (0,1)$ controls the balance of the contribution of the two terms in the distance and σ controls how beat distance deviations to τ_T affect $d(b_i, b_j)$.

3.3. Coping with Tempo Variations

It is desirable to adapt the inter-beat distances to the tempo variations that occur within a music excerpt. To do so, after the global tempo estimation process described in Section 2, a more precise tempo estimation process takes place. Firstly, a rough tempo curve $\mathbf{t}_c$ is generated by considering the most salient peak for each segment *s* around the found tempo *T*:

$$\mathbf{t}_c(s) = \underset{(1-\beta)T<t<(1+\beta)T}{\arg\max} \{\mathbf{TG}^{fl}(t,s) + \mathbf{TG}^{ch}(t,s)\} \quad (10)$$

for β value being around 0.1. Next, we re-estimate the tempo as described in Section 2 by using $Q = 4$ and all resonators with period resolution 5ms in tempo range of $[\min(\mathbf{t}_c), \max(\mathbf{t}_c)]$. Then the beat saliencies $\hat{s}_j^b$ are re-estimated as in Eq. (7) where

$$\hat{\mathbf{r}}_c(b_j) = \sum_i \mathbf{r}^i_{c,T(b_j)}(b_j) / \max_s\left(\sum_i \mathbf{r}^i_{c,T(s)}(s)\right),\ c = \{fl, ch\} \quad (11)$$

T(s) denotes the local tempo with the maximum strength at time segment *s* and $\mathbf{r}^i_{fl,T(s)}, \mathbf{r}^k_{ch,T(s)}$ denote the corresponding resonator responses. Then Eq. (9) is reformulated as

$$d_T(b_i, b_j) = 1 - \exp\{-\frac{1}{\sigma^2}\ln^2\left((b_j - b_i)/\tau_{T(b_j)}\right)\} \quad (12)$$

3.4. Dynamic Programming Solving

Let $\{b_l\}, l \in L \subseteq \{1..N\}$ be a target beat sequence. The optimal beat sequence $\{b_l^*\}$ should minimize the objective function

$$O(\{b_l^*, l \in L\}) = \sum_{l \in L} d(b_{l-1}^*, b_l^*) \quad (13)$$

We denote by $C^*(b_i)$ as the minimum cost to "reach" beat b_i. Establishing a dynamic programming schema, we write the recursive formulae

$$C^*(b_i) = \min_{b_k}\{d(b_k, b_i) + C^*(b_k)\} \quad (14)$$

$$path(b_i) = \arg\min_{b_k}\{d(b_k, b_i) + C^*(b_k)\} \quad (15)$$

where $path(b_i)$ denotes the preceding beat to reach b_i optimally. The costs $C^*(b_i), i = 1..N$ are calculated from Eq. (14) recursively. To derive the optimal sequence we choose a subset of possible beat locations $\{b_m'\}$ in the end of the excerpt. Similarly to Viterbi algorithm, the last beat is chosen as

$$b_K = \arg\min_{b_m'}\{C^*(b_m')\} \quad (16)$$

and the optimal beat sequence is found by moving backwards:

$$b_{l-1}^* = path(b_l^*),\ \ l = K..2 \quad (17)$$

4. EVALUATION

4.1. Tempo Estimation Results

The proposed tempo estimation method was ranked first in the MIREX 2011 Tempo Estimation task and outperformed all other submitted methods (Table 1). It was also evaluated on the ISMIR 2004 Tempo Induction evaluation exchange. Details on the database can be found in [7]. Table 2 shows the performance of the proposed method based on accuracies *acc1* and *acc2* within a 4% tolerance for three basic settings: metrical analysis with no source separation (MA); source separation (no metrical analysis), where tempo is decided by peak picking within $\mathbf{T}_{gl}$ (SS); both source separation and metrical analysis (SSMA). Each version was evaluated for all feature settings.

Results indicate that the deployment of the metrical analysis method presented, increases significantly *acc1*, in both datasets, and for all feature sets, except in the case of using solely filterbank energies in the ballroom dataset. For songs dataset, the usage of metrical relations increases *acc1* over 25% for all feature sets. For ballroom dataset, the improvement in *acc1* is smaller, but still significant

When combining features, the employment of source separation has different effect on the data-sets. *Acc1* and *acc2* increase by 1.7% and 0.8% respectively for the ballroom dataset. On the other hand, *acc1* decreases by 1.5% in songs dataset, while *acc2* increases by 0.8%. This can be explained by the fact that since many excerpts in Songs dataset are not percussive, the decomposition of a non percussive signal into a percussive and harmonic part will result to a meaningless and noisy decomposition. However, the source separation for tempo estimation task is promising and should be investigated in more detail in the future.

Table 3 shows comparative results of our method with the current state-of-the-art algorithms that reported results in ballroom/songs datasets, i.e. the top three performing algorithms at the ISMIR 2004 tempo induction contest [7], the baseline of our method [10] and Seyerlehner et al. [5]. It can be seen that the proposed method performance is at par within the current state-of-the-art algorithms.

It must be noted that the method proposed in [3] employs prior information about tempi, while the method presented in [11] is biased, since it was implicitly trained on these datasets.

4.2. Beat Estimation Results

The proposed beat tracking method was also submitted in MIREX 2011. Results are presented in Table 1. In MCK dataset, where tempo is almost constant, although the proposed method is ranked 6th it performs close to the best performing algorithms. On the other hand, in MAZ dataset where excerpts exhibit great tempo variations, the proposed method seems inadequate to capture beat locations. This is evident, since the algorithm searches for an almost constant tempo before accessing the beat locations.

Beat Tracking				**Tempo Estimation**	
MCK (F-Measure)		MAZ (F-Measure)		MCK (P-Score)	
SB3	52.69	FW1	67.56	**GKC3**	**0.8290**
SB4	50.86	SB4	51.17	FW2	0.7385
KFRO1	50.68	GP4	49.12	ZG1	0.7275
KFRO2	50.45	GP5	47.02	SP1	0.7105
GP5	50.32	**GKC2**	**42.18**	GKC6	0.6777
GKC2	**50.10**	GP2	41.80	SB5	0.6559
GP4	50.09	SB3	40.29		
GP3	49.56	GP3	40.16		

Table 1. MIREX 2011 beat tracking and tempo estimation algorithms (proposed method is denoted by GKC2 and GKC3).

		Ballroom		**Songs**	
	Features	**Acc1**	**Acc2**	**Acc1**	**Acc2**
MA	Filterbank	38.68	84.53	67.96	88.82
	Chroma	56.30	88.83	44.52	79.78
	Combined	58.17	92.41	60.00	89.03
SS	Filterbank	52.29	88.40	24.52	88.60
	Chroma	46.13	88.40	15.05	70.54
	Combined	53.30	90.97	21.08	88.17
SSMA	Filterbank	48.42	90.40	57.85	88.39
	Chroma	54.58	82.95	40.86	73.12
	Combined	59.89	93.27	58.49	89.89

Table 2. Effect of source separation and metrical analysis to tempo estimation accuracy.

	Ballroom		**Songs**	
	Acc1	**Acc2**	**Acc1**	**Acc2**
SSMA	59.89	93.27	58.49	89.89
MA	58.17	92.41	60.00	89.03
Klapupi [3,7]	63.18	90.97	58.49	91.18
Gkiokas [10]	61.08	93.98	42.15	90.11
Uhle [7]	56.45	81.09	41.94	71.83
Scheirer [4,7]	51.86	75.07	37.85	69.46
SE1 [11]	78.51	-	40.86	-
SE2 [11]	73.78	-	60.43	-

Table 3. Comparative results of the proposed method to the state-of-the-art.

5. CONCLUSION AND FURTHER WORK

In this paper we presented Tempo Estimation and Beat Tracking algorithms. Although no prior information is assumed, both algorithms perform within the current state-of-the-art. Simple metrical analysis enhances greatly the selection of the correct tempo compared to peak-picking of the periodicity vector. This analysis may be extended to take into account more metrical relations. Harmonic/Percussive separation, which has not been used for tempo estimation and beat tracking before, seems to enhance the accuracy of tempo estimation in some cases. However, the utilization of harmonic/percussive separation should be investigated more thoroughly in the future, since both parts of the music signal contain complementary and sometimes conflicting rhythmic information.

6. REFERENCES

[1] Davies M., Plumbley M., "Context-Dependent Beat Tracking of Musical Audio", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, March 2007.

[2] Ellis D. "Beat Tracking by Dynamic Programming", *J. New Music Research*, 26(1):51–60, 2007.

[3] Klapuri A., Eronen A. and Astola J., "Analysis of the Meter of Music Acoustic Signals", *IEEE - Transactions on Audio, Speech and Language Processing*, 14(1), January 2006.

[4] Scheirer E., "Tempo and Beat Analysis of Acoustic Musical Signals.", *The Journal of the Acoustical Society of America*, Vol. 103, No. 1, January 1998

[5] S. Dixon. "Automatic extraction of tempo and beat from expressive performances. *J. New Music Research*, 30(1):39–58, 2001.

[6] Peeters G., "Rhythm Classification Using Spectral Rhythm Patterns", *Proceedings of the 6th International Conference on Music Information Retrieval*, London, UK, 2005

[7] Gouyon F., Klapuri A., Dixon S., Alonso M., Tzanetakis G., Uhle C., and Cano P., "An Experimental Comparison of Audio Tempo Induction Algorithms", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, September 2006.

[8] Alonso M., Richard G., David B., "Accurate Tempo Estimation Based on Harmonic + Noise Decomposition", *EURASIP Journal on Applied Signal Processing* Volume 2007, Issue 1, January 2007

[9] Eronen A., Klapuri A., "Music Tempo Estimation with k-NN Regression", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18 No. 1, January 2010.

[10] Gkiokas A., Katsouros V. and Carayiannis G., "Tempo Induction Using Filterbank Analysis and Tonal Features", *Proceedings of the 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, August 2010.

[11] Seyerlehner K., Widmer G., and Schnitzer D., "From Rhythm Patterns to Perceived Tempo", *Proceedings of the 8th International Conference on Music Information Retrieval,* Vienna, Austria, 2007

[12] Eyben F., Böck S., Schuller B. and Graves A. "Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks", *Proceedings of the 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, August 2010.

[13] Peeters G. and Papadopoulos H., "Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation", *IEEE - Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 6, August 2011.

[14] FitzGerald D. "Harmonic/Percussive Separation Using Median Filtering", *Proceedings of the 13th International Conference on Digital Audio Effects*, Graz, Austria, 2010.