ILSP AUDIO BEAT TRACKING ALGORITHM FOR MIREX 2012

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

{agkiokas, vsk}@ilsp.gr, gcara@ilsp.athena-innovation.gr

[1] Institute for Language and Speech Processing / "R.C Athena"
[2] National Technical University of Athens

ABSTRACT

This paper describes a beat tracking system submitted to the MIREX 2012 and is described in detail in [1]. Two main feature classes are extracted 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. Beat tracking involves the computation of the beat saliencies derived from the resonators responses and defines a distance measure between candidate beats locations. A dynamic programming algorithm is adopted to the find the optimal "path" of beats.

1. FEATURE EXTRACTION

1.1 Pre-analysis

The constant Q transform (CQT) of the audio signal is calculated on the whole input signal, using 12 bins per octave, with 25Hz and 5kHz minimum/maximum frequencies respectively (Q value equals to 17), and a Hanning window with half overlap. Frequency bins are aligned to the western scale musical pitches. The frequency bins are rescaled by bicubic interpolation/decimation to have equal frames per time unit (200 frames/s), resulting the log-frequency spectrogram $\mathbf{S} = \{S_{i,f}\}$ where i and f denote the time and frequency bin indices respectively.

1.2 Chroma and Filterbank Energies

The percussive/harmonic separation algorithm presented in [2] is applied to the CQT of the signal. Chroma vectors and the energies of 8 triangular filters in the mel scale are calculated from the harmonic/percussive part of the signal respectively. We denote filterbank energies as $\mathbf{x}_{fl}$ and chroma features as $\mathbf{x}_{ch}$.

2. PERIODICITY ANALYSIS

Feature vectors are differentiated and convolved with a bank of resonators as in [2] in the range of $[30, 500]$ bmp, resulting $\mathbf{TG}^{fl}$ and $\mathbf{TG}^{ch}$ periodicity vectors for filterbank energies and chroma features respectively. To estimate the global periodicity vector $\mathbf{T}_{gl}$ for the whole excerpt $\mathbf{TG}^{fl}$ and $\mathbf{TG}^{ch}$ are summed across all segments and then multiplied:

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

3. TEMPO ESTIMATION

We compute the fundamental tempo T_0 as

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

Then we expect that $\mathbf{T}_{gl}$ has peaks at target tempo as well as at integer multiples of T_0. We consider a model of two tempi $\{T_{slow}, T_{fast}\}$ values under the assumption that T_{slow} is the more perceptually relevant, while T_{fast} is more likely to be double, triple or quadruple of T_{slow}.

We define the joint salience J_s of T_{slow}, T_{fast} as

$$J_s(T_{slow}, T_{fast}) = [T_{gl}(T_{slow}) + T_{gl}(T_{fast})] \cdot \sum_{i=2..4} e^{-(\frac{T_{fast}}{T_{slow}} - i)^2 / (\sigma i)^2} \qquad (2)$$

The final tempo T is the T_{slow} that maximizes J_s and is multiple of T_0:

$$T = \arg\max_{iT_0} \{J_s(iT_0, kT_0), iT_0, kT_0 \in \{30,..500\}\} \qquad (3)$$

4. BEAT TRACKING

As beat candidates we consider the peaks of the resonator responses corresponding to the found tempo T for all feature sequences $\mathbf{x}_{ch}^{i}$, $\mathbf{x}_{fl}^{k}$. We denote the time instances of beat candidates as $\{b_j\}, j = 1..N$ and $\mathbf{r}_{fl}^{k}$, $\mathbf{r}_{ch}^{j}$ the re-

sponses of the resonator for feature sequences $\mathbf{x}_{fl}^{k}$ and $\mathbf{x}_{ch}^{i}$ respectively. The saliencies of beat candidates $\{b_j\}$, denoted as s_j^b and are computed as:

$$s_j^b = \hat{\mathbf{r}}_{fl}(b_j) + \hat{\mathbf{r}}_{fl}(b_j) \quad (4)$$

where

$$\hat{\mathbf{r}}_{type}(k) = \sum_i \mathbf{r}_{type}^i(k) / \max_s \left(\sum_i \mathbf{r}_{type}^i(s) \right), \ type = \{fl, ch\} \quad (5)$$

4.1 Inter-Beat Distances

The distances between b_i and b_j, $b_j > b_i$ denoted by $d(b_{i,}b_j)$ as

$$d(b_i, b_j) = \gamma \cdot d_t(b_i, b_j) - (1-\gamma) s_j^b \quad (6)$$

where

$$d_t(b_i, b_j) = 1 - \exp\{-\frac{1}{\sigma^2}\left(\ln\left(\frac{b_j - b_i}{T} \right)\right)^2\} \quad (7)$$

4.2 Coping with Tempo Variations

After the global tempo estimation process described in Section 2, a more precise tempo estimation process in both time and frequency space takes place. Firstly, a rough tempo curve $\mathbf{t}_c$ is generated from $\mathbf{TG}^{fl} + \mathbf{TG}^{ch}$ by considering the most salient peak for each segment s around the found tempo T:

$$\mathbf{t}_c(s) = \underset{(1-\beta)T<t<(1+\beta)T}{\arg\max} \{\mathbf{TG}^{fl}(t,s) + \mathbf{TG}^{ch}(t,s)\} \quad (8)$$

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

$$\hat{\mathbf{r}}_{type}(b_j) = \sum_i \mathbf{r}_{type,T(b_j)}^i(b_j) / \max_s \left(\sum_i \mathbf{r}_{type,T(s)}^i(s) \right), \ type = \{fl, ch\} \quad (9)$$

$T(s)$ denotes the local tempo with the maximum strength at time segment s and $\mathbf{r}_{fl,T(s)}^i, \mathbf{r}_{ch,T(s)}^k$ denote the corresponding resonator responses.

4.3 Dynamic Programming Solving

Let $\{b_l\}, l \in L \subseteq \{1..N\}$ be a target beat sequence. Then according to Eq. 9 the optimal beat sequence $\{b_l^*\}$ should minimize the objective function

$$O(\{b_l^*, l \in L\}) = \sum_{l \in L} d(b_{l-1}^*, b_l^*) \quad (10)$$

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

$$C^*(b_i) = \min_{b_k} \{d(b_k, b_i) + C^*(b_k)\} \quad (11)$$

$$path(b_i) = \arg\min_{b_k} \{d(b_k, b_i) + C^*(b_k)\} \quad (12)$$

where $path(b_i)$ denotes the preceding beat to reach b_i optimally. We estimate the optimal sequence by

$$b_K = \arg\min_{b_m^{'}} \{C^*(b_m^{'})\} \quad (13)$$

and the optimal beat sequence is found by moving backwards:

$$b_{l-1}^* = path(b_l), \ l = K..2 \quad (14)$$

5. REFERENCES

[1] Gkiokas A., Katsouros V., Carayannis G. and Stafylakis T., "Music Tempo Estimation and Beat Tracking by Applying Source Separation and Metrical Relations," in *Proc. of the 37th IEEE ICASSP*, Kyoto, Japan, March 25-30, 2012.

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