ILSP AUDIO TEMPO ESTIMATION ALGORITHM FOR MIREX 2011

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 tempo estimation system submitted to the MIREX 2011 and is an extension of the work presented in [2]. 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.

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. From the harmonic/percussive part the chroma vectors and the energies of 8 triangular filters in the mel scale are calculated respectively.

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_1 is the T_{slow} that maximizes J_s and is multiple of T_0:

$$T_1 = \arg\max_{iT_0} \{J_s(iT_0, kT_0), iT_0, kT_0 \in \{30, ..500\}\} \qquad (2)$$

If $T_1 = T_0$ then tempo T_2 is chosen equal to T_{fast} otherwise $T_2 = T_0$.

4. 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.