A Tempo Extraction Algorithm for Raw Audio Recordings

Iasonas Antonopoulos, Aggelos Pikrakis and Sergios Theodoridis
Department of Informatics and Telecommunications
University of Athens
Panepistimioupolis, 15784, Athens, Greece
{jantonop, pikrakis, stheodor}@di.uoa.gr

Abstract

This paper presents a tempo extraction algorithm for raw, polyphonic audio recordings, assuming that music meter and tempo remain constant throughout the recording. The method was submitted at MIREX 2006, in the context of the Audio Tempo Extraction evaluation task. Our approach is based on the self-similarity analysis of the audio recording and does not assume the presence of percussive instruments. In order to account for the contest's requirements, which emphasizes on perceptual tempo tracking, the proposed method returns two tempi, ranging from 40bpm to 320bpm, along with their relative strength.

Keywords: Tempo Extraction, Self Similarity Analysis

1. Description of the Algorithm

1.1. Feature Extraction

The proposed algorithm is a variant of previously published work by the authors ([1]). At a first step, each raw audio recording is divided into overlapping long-term segments, each of which has a duration equal to 6 seconds (long-term step has been set equal to 0.5 seconds). For each long-term segment, a short-term moving window generates a sequence of feature vectors. Approximate values for the length of the short term window and overlap between successive windows are 100ms and 95ms respectively. The adopted feature is similar with the standard MFCCs, however the filter bank consists of overlapping triangular filters, whose center frequencies align with the chromatic scale of tones (starting from 110Hz and reaching up to approximately 7KHz).

Let $\mathbf{F} = \{f_1, f_2, \ldots, f_N\}$, be the feature sequence that is extracted from a long-term segment. Sequence F serves as the basis to calculate the Self Similarity Matrix (SSM) of the segment, using the Euclidean function as a distance metric. Since the SSM is symmetric around the main diagonal, in the sequel it suffices to focus on its lower triangle.

At a next step, the mean value of each diagonal of the SSM is calculated. If B_k stands for the mean value of the kth diagonal, then:

$$B_k = \frac{1}{N-k}\sum_{l=k}^{N} ||f_l, f_{l-k}||$$

where $N - k$ is the length of the k-th diagonal and $||.||$ is the Euclidean distance function. In the sequel, we will refer to the k-th diagonal as the k-th lag. If B is treated as a function of k, then its plot against k exhibits certain local minima (valleys) for a number of ks. Each valley can be interpreted as corresponding to a periodicity, that is inherent in the long-term segment being examined. In addition, the difference of lags between any two valleys can also reveal inherent periodicities.

1.2. Tempo Estimation

Our method is based on the assumption that the perceived tempi correspond to periodicities that appear as valleys of B. In order to extract a pair of dominant periodicities (i.e., two perceived tempi), each long-term segment is processed as follows:

1. All "dominant" valleys of B are detected. A valley is considered to be "dominant" if it is the deepest one in a neighborhood of lags. The width of the neighborhood is a predefined constant. The lags of detected valleys are sorted in ascending order and the difference of lags between successive valleys is calculated. All differences are placed in a histogram. This histogram exhibits peaks at certain values.

2. In the sequel, the width of the neighborhood for calculating dominance is increased and all dominant valleys are again detected for this new width. The differences of lags of detected valleys are placed in a new histogram.

3. The peaks of the above two histograms are then detected and all possible pairs of lags corresponding to these peaks are formed. The ratio of lags in each pair is then examined in order to decide whether it approximates sufficiently one of the following music meter ratios, namely: $\{\frac{3}{8}, \frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}, \frac{8}{8}, \frac{9}{8}\}$, or $\{\frac{2}{4}, \frac{3}{4}, \frac{4}{4}, \frac{5}{4}\}$, depending on the value of the lag of the first component of the pair under consideration. The best pair is selected by also taking into account the height of

peaks corresponding to the lags of the pair in the respective histograms.

The above procedure yields one pair of lags per long-term segment. At a final stage the lags appearing in all winner-pairs are again placed in a histogram. This last histogram exhibits certain peaks. These peaks are examined in pairs and the procedure of step 3 is repeated to yield a winning pair. The two lags of this pair are returned by the algorithm as the two most dominant periodicities (perceived tempi).

The proposed method took the 6th place in the MIREX-2006 tempo extraction contest. The performance of the algorithm is summarized in Table 1. The method was implemented in Matlab (v7.0) and was cross-compiled to generate the submitted executable.

At least one tempo correct	**Both tempi correct**	**P-score**
84.29% (5th place)	47.86% (3d place)	0.669 (6th place)

Table 1. Performance of the proposed method

References

[1] Aggelos Pikrakis, Iasonas Antonopoulos and Sergios Theodoridis, “Music Meter and Tempo Tracking from raw polyphonic audio”, in *Proc. of ISMIR 2004*, Barcelona, Spain, Oct. 2004.