TEMPO ESTIMATION BY MODELLING PERCEPTUAL SPEED

Anders Elowsson **Anders Friberg**
KTH Royal Institute of Technology
elov@kth.se afriberg@kth.se

ABSTRACT

This paper describes a tempo estimation system submitted to MIREX 2013. The proposed algorithm detects onsets from harmonic/percussive-separated audio. Perceptual tempo is modeled in a multiple linear regression as a function of onset densities, difference in sound level of the harmonic and percussive part, IOIs of clustered components, as well as spectral fluctuations. Periodicity histograms are generated based on IOIs between onsets, where the contribution to the histogram of each IOI is weighted to account for spectral properties of the onsets as well as salience. Finally peaks of the histograms are chosen in a logistic regression model based on peak height and estimated perceptual speed.

1. INTRODUCTION

The tempo estimation system TEMPS (**T**empo **E**stimation by **M**odeling **P**erceptual **S**peed) uses the perceptual speed of music to overcome the well-known octave error. A flowchart of the program is shown in Figure 1.

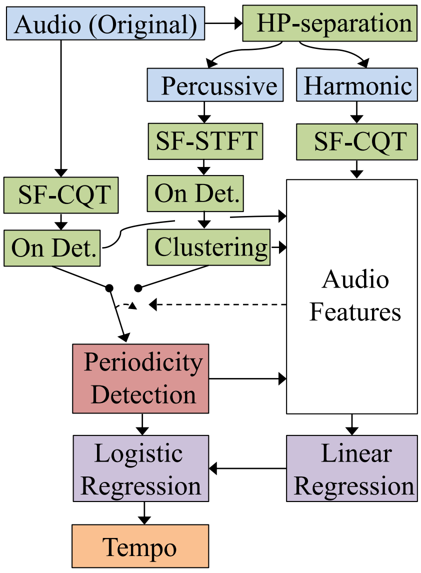

Figure 1. Flowchart of the processes used to compute the tempo of music audio.

In earlier works a computational model of speed in music audio has been developed using a custom set of rhythmic features [1]. The test set of the MIREX challenge is not available to submitters and consists of 140 songs of varying genres, tempo and metre. A training set was created consisting of over 800 songs (most taken from the Ballroom and Songs dataset). The original tempo annotations could not be used as they did not correspond to perceptual tempo. Therefore the first author annotated the songs with two tempi and a relative weight (*w*) between the two in accordance with the MIREX tempo estimation task.

2. SOURCE SEPARATION AND ONSET DETECTION

The first step of the system is to separate harmonic and percussive content, similar to [5], based on the method proposed in [3], with parameters described in [1]. Onsets are detected by peak picking in a half-wave rectified spectral flux. Percussive onsets are detected from the percussive waveform using a STFT and harmonic onsets are detected with a CQT [4] as described in [1]. The percussive onsets are clustered into different components with a K-means clustering as in [2]. To determine if the percussive onsets or the harmonic onsets should be sent to the periodicity detection, features such as sound level in the harmonic and percussive waveform are used.

3. PERIODICITY

To detect periodicity, a histogram over onset distances is generated (*periodogram*), where the contribution of each onset-pair increases with increasing similarity in spectrum as well as increasing onset strength. Often there are several peaks that may correspond to the correct period, as shown in Figure 2.

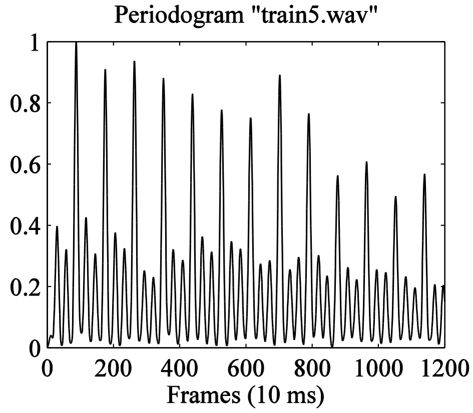

Figure 2. A periodogram, computed as the histogram of onset distances, weighted so that the contribution of each onset-pair increases with increasing similarity in spectrum as well as increased salience.

To find the most salient period in this case, the FFT of the periodogram is computed. The resulting spectrogram of the periodogram will be referred to as the *cepstroid*. After converting the cepstroid back to the time-domain it is convolved with the periodicity vector and the highest peak is chosen as the period as shown in Figure 3.

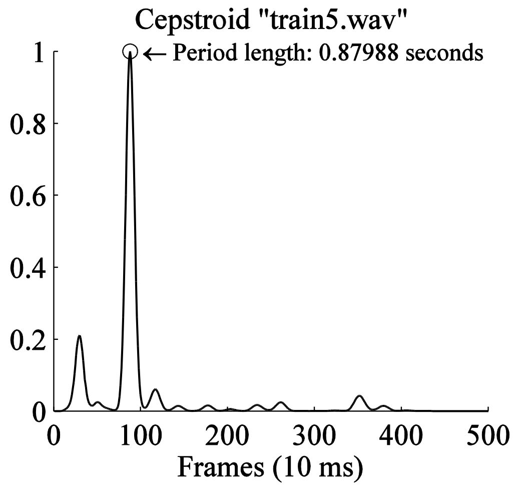

Figure 3. The cepstroid, computed as the spectrum of the periodogram.

4. REGRESSION MODELS

First, perceptual speed of the music is modelled in a multiple linear regression as a function of onset densities, difference in sound level of the harmonic and percussive part, IOIs of clustered components, as well as spectral fluctuations. As there were no speed annotations in the training data, ground truth speed was approximated with the tempo annotations by

$$\text{Speed} = \log(\text{T}_1)w + \log(\text{T}_2)(1-w)$$

Figure 4 shows the modelling of “speed” for 586 songs.

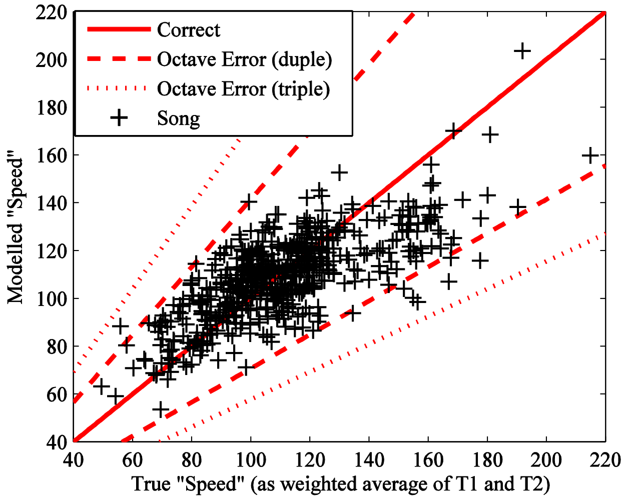

Figure 4. The modelling of “speed” for 586 percussive songs in the training set. The dashed line indicates the position where an octave error would be produced in duple metre and the dotted line indicates octave error for triple metre.

As the final step, beat lengths are evaluated at positions corresponding to

$$P_{len} \times \left(\frac{1}{2}\right)^{n}, \quad P_{len} \times \left(\frac{1}{2}\right)^{n} \times \left(\frac{1}{3}\right) \qquad n = 0,1,2,..$$

where P_{len} is the period length found in the periodicity detection step. Features are calculated for each peak position based on ceptroid height, periodogram height, beatogram height (same as periodogram but the contribution of each onset-pair increases with increasing *dissimilarity* in spectrum), distance from perceptual speed, and the peak’s ratio of the period length. The features are evaluated in a logistic regression.

5. RESULTS

The reported P-Score of the system was about 0.86. As shown in Figure 5 below, this is the highest results reported so far in the competition, among 35 contributions.

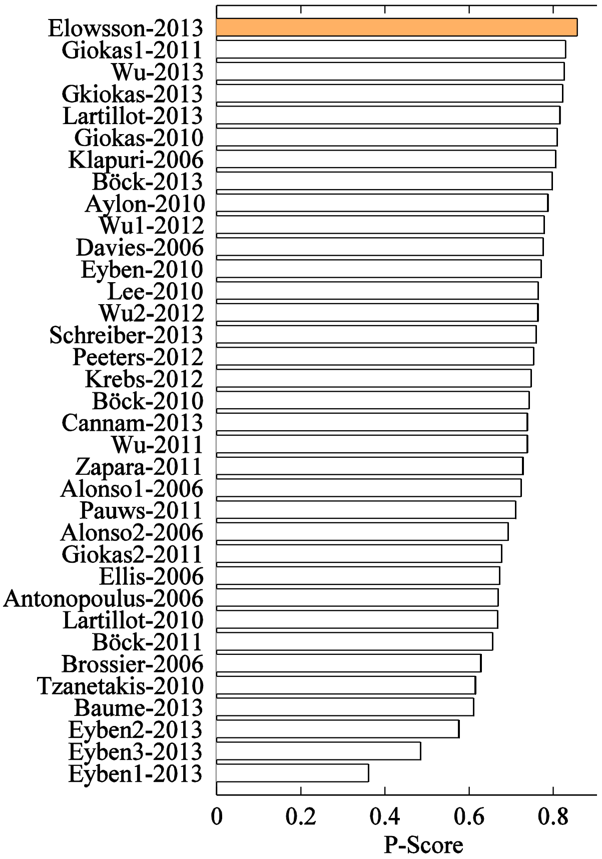

Figure 5. The results for all 35 contributions to the MIREX tempo estimation task, since the present P-Score measurement was established in 2006. The proposed system (orange bar at the top) achieves the highest result so far.

6. CONCLUSIONS

The good results of the system indicate that the proposed method accurately models tempo. The novel approach, with two regression models connected in series captures two different dimensions relevant to perceived tempo.

With a linear regression we model the speed of the music, and in a logistic regression framework we model pulse strength. A third dimension, metre, will be addressed in future work. For this dimension we expect the periodogram and the cepstroid to be important.

The P-Score for the training set was about 0.97 in a leave-one-out cross-validation, indicating that even better results are plausible with training examples more consistent with the test set or with better ground truth annotations of the training set.

7. REFERENCES

[1] A, Elowsson and A. Friberg: "Modelling the Speed of Music Using Features from Harmonic/Percussive Separated Audio," In Proc. of ISMIR, pp. 481-486, 2013.

[2] A, Elowsson and A. Friberg: "Modelling Perception of Speed in Music Audio." In Proc. of the Sound and Music Computing Conference 2013," pp. 735-741, 2013.

[3] D. FitzGerald: "Harmonic/Percussive Separation Using Median Filtering," In Proc. of DAFx, 2010.

[4] C. Schörkhuber and A. Klapuri: "Constant-Q Transform Toolbox for Music Processing," In 7th Sound and Music Conference, Barcelona, 2010.

[5] A. Gkiokas, V. Katsouros and G. Carayannis: "Reducing Tempo Octave Errors by Periodicity Vector Coding And SVM Learning," In Proc. of ISMIR, pp. 301-306, 2012.