A SEGMENTATION-BASED TEMPO INDUCTION METHOD

Maxime Le Coz, Helene Lachambre, Lionel Koenig and Regine Andre-Obrecht
IRIT, Universite Paul Sabatier,
118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9
{lecoz,lachambre,koenig,obrecht}@irit.fr

ABSTRACT

The automatized beat detection and localization have been the subject of multiple research in the field of music information retrieval. Most of the methods are based on onset detection. We propose an alternative approach:

Our method is based on the "Forward-Backward segmentation": the segments may be interpreted as attacks, decays, sustains and releases of notes. We process the segment boundaries as a weighted Dirac signal. Three methods devived from its spectral analysis are proposed to find a periodicity which corresponds to the tempo.

The experiments are carried out on a corpus of 100 songs of the RWC database. The performances of our system on this base demonstrate a potential in the use of a " Forward-Backward Segmentation" for temporal information retrieval in musical signals.

1. INTRODUCTION

The automatized beat detection and localization have been the subject of multiple research in the field of music information retrieval. The study of beat is indeed important as the structure of a music piece lies in the beat. Western music uses however different levels in the hierarchy of scale measuring time. We have to distinguish the *tatum* which is "the regular time division that mostly coincides with all note onsets" [3] from the *tactus* which is defined as the rate at which most people would clap their hands when listening to the music [8]. Here, we look for the *tactus*, which will be named tempo and measured in beat per minute (BPM).

Several methods have been suggested in order to extract the tempo information from an audio signal. Most of them use an onset detection method as onset localization carries the temporal structure that leads to the estimation of the tempo. Theses methods use different observation features in order to propose a list of onset positions. They are very dependent on that detection. Dixon's first algoritm [4] uses an energy based detector in order to track the onset positions. Then a clustering is performed on the inter-onset-interval values. Some best clusters are chosen as possible hypothesis. A hypothesis is finally validated with a beat tracking.

In Alonso's algorithm [1], onset positions are deducted by using a time-frequency representation and a differentiator FIR filter to detect sudden changes in the dynamics, timbre or harmonic structure. The tempo is then deduced using either the autocorrelation or spectral product.

Klapuri [9] proposes a more complex way of extracting the onset positions. The loudness differentials in frequency subbands are computed and combined in order to create four accent bands. This aims at detecting harmonic or melodic changes as well as percussive changes. Using comb filter resonators to extract features, and probalistic models, the values of *tatum*, *tactus* and *measure* meter are computed.

Uhle [12] suggests a method based on the segmentation of the signal into long-term segments corresponding to its musical structure (for example, the verses and chorus of a song). The amplitude envelope of logarithmically spaced frequency subbands is computed; its slope signal aims to represent accentuation on the signal. The analysis of an autocorrelation function on 2.5 second segments inside each long-term segment gives the tatum estimator. A larger-scale analysis over 7.5 second segments is then performed in order to give values corresponding to the measure. The local maxima positions of the autocorrelation function are finally compared with a bank of pre-defined patterns in order to define the best value of the tempo on the long term segment.

Dixon [5] has proposed an alternative method to onset calculation. The signal is splitted into 8 frequency bands and autocorrelation is performed on each smoothed and downsampled subband. The three highest peaks of each band are selected and combined in order to determine the final tempo estimation.

Another algorithm is that of Scheirer [10]. This algorithm performs a comb filterbank that seeks for periodically spaced clock pulse that best matches the envelope of 6 frequency subbands.

Tzanetakis [11] suggests a method based on a wavelet transform analysis. This analysis is performed over 3 second signal segments with 50% of overlap. On each segment, amplitude envelope of 5 octave-spaced frequency bands is extracted. Autocorrelation is then computed. Three

kind of autocorrelation analysis are computed in order to estimate the value of the tempo. The first one is the median of highest peak of the sum of the envelopes over every window. The second one returns the median value of the highest peak on each subband and each segment. The last one computes several best peaks from the autocorrelation on the sum of every envelope and then chooses the most frequent value.

Our method is based on the analysis of an automatic segmentation of the signal into quasi-stationary segments : the segments may be interpreted as attacks, decays, sustains and releases of notes. So we propose to process the segment boundaries in order to find a periodicity which would correspond to the tempo.

In section 2, we describe the segmentation used as a front-end, the analysis of this segmentation in the frequency domain and the different methods we use to extract the value of the tempo in BPM. In the last part, we present the results of our experiments on the RWC [6, 7] corpus.

2. METHOD

Our method relies on the detection of quasi-stationnary segments in the audio signal waveform. A frequency analysis of the boundaries is then performed in order to find the most present periodicities and thereby estimate the tempo consequently.

The algorithm is based on three steps :

- Segmentation
- Boundary frequencial analysis
- Tempo extraction

2.1 Segmentation

We segment the signal using the "Forward Backward Divergence" [2]. The signal is assumed to be a sequence of quasi-stationnary units, each one characterized by the following gaussian autoregressive model :

$$\begin{cases} y_n = \sum a_i y_{n-i} + e_n \\ var(e_n) = \sigma_n^2 \end{cases} \quad (1)$$

where y_n is the signal and e_n an uncorrelated zero mean Gaussian sequence.
As the variance σ_n is constant over an unit and equals σ, the model of each area is parametered by the following vector :

$$(A^T, \sigma) = (a_1, ..., a_p, \sigma) \quad (2)$$

The strategy is to detect changes in the parameters, using a distance based on the mutual conditional entropy. A subjective analysis of the segmentation shows a sub note segmentation and the location of attacks, sustains and releases.

For a solo musical sound, the segments of the signal correspond to the different steps of a note. On Figure 1, we present a solo note of trombone. The note is segmented into four parts, which correspond to the attack, the sustain and the release. Note that the attack and decay phases of some notes are often grouped together into a single segment. In such cases, the attack period is too short for the segmentation algorithm as it imposes a minimal length to initiate the autoregressive model.

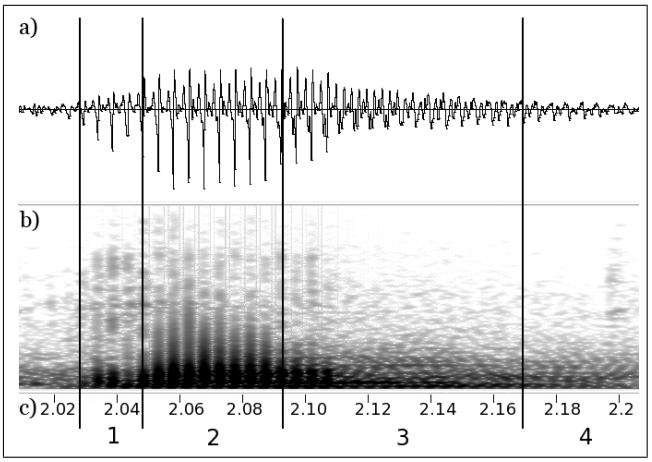

Figure 1. Segmentation of a trombone note. a) Waveform, b) Spectrogram, c) Time. 1) Attack, 2) Sustain, 3 & 4) Release. The vertical lines are the boundaries of the segments. The first boundary correspond to the onset.

As they represent a rupture point of the signal, we assume that onset localizations, containing the tempo information, are included in the list of boundaries time. We therefore focus on positions of the boundaries.

2.2 Boundary Frequencial analysis

The main objective is to find a periodicity in the localization of the boundaries that would be the effect of the song's rythmical pattern. In order to find the periodicity, a signal $b_w(t)$ is created. This signal is a weighted Dirac signal, where each Dirac is positioned at the time of a boundary t_k.

The Diracs are weighted in order to give more influence to the boundaries located at times that are most likely to be onsets. Asuming that at onset times, an increase of energy is observed, each Dirac is weighted by the difference between the energy of the spectrum computed on 20 ms after and before t_k (resp. e_k^+ and e_k^-).

$$w(t_k) = e_k^+ - e_k^- \quad (3)$$

We obtain $b_w(t)$ (see an example on Figure 2) :

$$b_w(t) = \sum_{k=1}^{N} \delta(t - t_k) w(t_k) \quad (4)$$

where N is the count of boundaries, t_k is the time of the k^{th} boundary.

We compute B_w, the Fourier transform of b_w to extract frequency information of this signal.

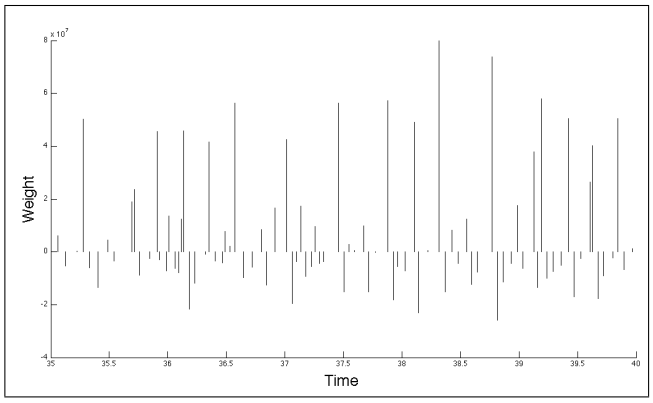

Figure 2. Representation of a $b_w(t)$

The expression of the Fourier transform $B_w(f)$ is :

$$\begin{aligned} B_w(f) &= \int_{\mathbb{R}} \sum_{k=1}^{N} \delta(t - t_k) e^{-2i\pi ft} w(t_k) dt \\ &= \sum_{k=1}^{N} e^{-2i\pi f t_k} w(t_k) \end{aligned} \tag{5}$$

This formula offers the advantage of being fast to calculate.

2.3 Tempo extraction

2.3.1 Spectrum analysis

We analyse the spectrum B_w on the range of frequencies 30 - 400 BPM (an example is given on Figure 3). We find the positions of the highest peaks as a base for each decision.

We then extract the positions and energies of the main peaks in terms of energy. As it is computed over a long time, the peaks of the spectrum are high and narrow, which makes the localization easier.

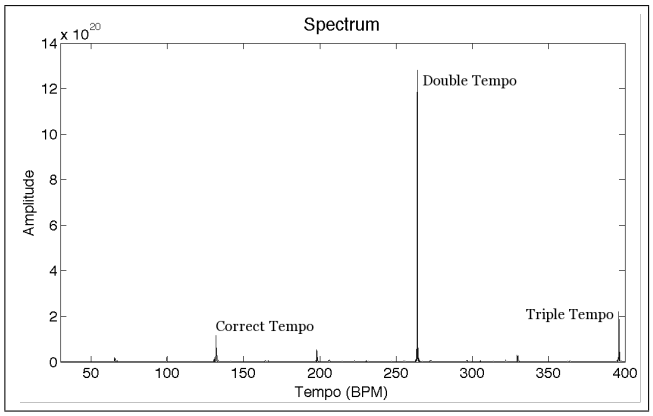

Figure 3. Spectrum $|(B_w(f)|^2$ of a whole song.

This localization is obtained by detecting the local maxima. This algorithm considers a point p and its two direct neighbors. p is a local maxima if

$$\begin{aligned} |B_w(p-1)|^2 < |B_w(p)|^2 \\ |B_w(p+1)|^2 < |B_w(p)|^2 \end{aligned} \tag{6}$$

We then choose several of the highest peaks with the only constraint that the distance between two peaks has to be greater than 3 BPM. Only a few peaks are really higher than others in the spectrum, so we choose to select only the four greatest peaks in terms of energy, the position selected for further peaks would be considered as noise. Let $P = \{p_1\ p_2\ p_3\ p_4\}$ be the list of selected peak positions under the constraint : $|B_w(p_i)|^2 > |B_w(p_{i+1})|^2$. We observe that every selected peak carries information that can be exploited in order to find out the value of the correct tempo. We finally apply a decision algorithm on P to find the tempo.

Two strategies are concidered. The first one looks for the correlation between the length of the segments and each value p in the temporal domain. The second one tries to find the best comb matching the spectrum.

2.3.2 "Inter-Boundaries-Intervals decision"

The first approach is in the temporal domain, and uses the boundaries of the segmentation. Theses boundaries are filtered on their weights in order to keep only the boundaries where a high increase of energy is experienced: we only keep the boundaries with a significant weight. This filtering is computed in order to keep instants which are most likely onset instants. The set I of intervals between each couple of following boundaries is then computed.

For each p_i, we perform the pseudo periods corresponding to 1/4, 1/3, 1/2, 1, 2, and 3 times p_i. These pseudo periods have been chosen as they correspond to the period of half, quarter, eighth and sixteenth note in duple meter or triple meter.

The score $Num(p_i)$ is the number of intervals in I whose durations correspond to one of these pseudo periods.

The estimated tempo $\widehat{p_b}$ is given by :

$$\widehat{p_b} = \underset{p_i, i=1,...,4}{\operatorname{argmax}} (Num(p_i)) \tag{7}$$

2.3.3 "Comb decision"

The second method uses the spectrum and is in frequency domain. This method is based on the first peak p_1, as we assume that it is always significant for the tempo detection. We then consider 7 tempi, which are $\frac{1}{4}p_1$, $\frac{1}{3}p_1$, $\frac{1}{2}p_1$, $2p_1$, $3p_1$ and $4p_1$, as well as p_1 itself, noted tp_i, $i = 1, ..., 7$. We only keep, among this list of tempi, those which are in the range 30 - 240 BPM, assuming that a value outside of these bounds would hardly be considered as the main tempo.

For each tempo value tp_i, we compute the product of the spectrum and a Dirac comb with the 10 harmonic teeth corresponding to the tempo value.

The mean amplitude value of the so filtred spectrum gives a score $Ampl(tp_i)$.

The estimated tempo $\widehat{p_c}$ is given by

$$\widehat{p_c} = \underset{tp_i, i=1,...,7}{\operatorname{argmax}} (Ampl(tp_i)) \tag{8}$$

2.3.4 Combination of the strategies

In order to take advantage of both methods, we propose a combined decision algorithm. Using p_{c1} and p_{c2} the two best tempi returned by the "Comb decision" algorithm, we apply the "Inter-Boundaries-Intervals" strategy to compare the two values $Num(p_{c1})$ and $Num(p_{c2})$.

The tempo with the best Num is chosen as a final decision.

3. EXPERIENCE

3.1 Corpus

We choose to test our method on the part of the RWC database [6, 7] that is BPM-annotated. This corpus has been created in order to provide a benchmark for experimentation on music information retrieval and is now well known and widely used in this research field. It therefore seems interesting to use it in order to facilitate comparisons between our algorithm's results and others. This corpus is a compilation of 100 tracks of Japanese Pop songs. Each song lasts from 2 minutes 50 seconds to 6 minutes 07 seconds.
As the method needs no learning, our experiment protocol consists in applying our algorithm on each full track.

3.2 Experiments

The methods are based on the Forward Backward divergence segmentation: in order to implement this algorithm, we choose to use the parameters defined in [2] for voiced speech signal. No specific adaptation is performed for music.

As previously mentioned, we observe that the highest peak of the spectrums has a strong link with the tempi. Over the 100 tracks computed, the highest peak position is linked with the tempo 98 times: it is located twice on a position corresponding to the half of the ground-truth tempo, 3 times on the correct position, 60 times on the double tempo and 32 times on a position corresponding to 4 times the tempo.

To assess quantitatively each version of our method, we introduce a confident interval : the tempo value is considered as "Correct" if its difference with the ground-truth value at strictly less than 4 BPM. The ratios and multiples are considered good when their distance to 2, 3, 4, 1/3 or 1/2 is strictly less than 0.03.

Two metrics are computed in order to evaluate the accuracy of each method. The first one is the ratio of correctly estimated tempi over the whole corpus.

$$Accuracy_1 = \frac{\#\ of\ correctly\ estimated\ tempi}{L} \quad (9)$$

where L is the number of evaluated tracks.

The second one is more flexible and assumes that the tempi corresponding to half, third double and three time the annotated tempo are correct. This metric is computed taking take into account that tempo value is subjective and can vary from one listener to another.

$$Accuracy_2 = \frac{\#\ of\ correct\ or\ multiple\ tempi}{L} \quad (10)$$

3.2.1 "Inter-Boundaries-Intervals decision"

The filltring of the boundaries involves a threshold: the selectionned boundaries have a weight greater than 10% of the maximum weight among the boundaries. The detailled results of the Inter-Boundaries-Intervals decision are visible in Table 1. The global result are 56 % of $Accuracy1$ and 95% of $Accuracy2$.

Ratios with the correct tempo						
1/2	1	2	4	No link	Acc_1	Acc_2
7	56	28	1	5	56	95

Table 1. "Inter-Boundaries decision Decision" : Number of music tracks in function of the ratios between the estimated tempo and the ground truth value. $Accuracy_1$ and $Accuracy_2$ are deducted.

3.2.2 Comb decision

In order to optimize the results of this method and to be sure to get the peak value on each hypothesis multiple, the returned value is the maximum of 7 equally spaced tempi in a neighborhood of $\pm 1\ BPM$ around each p multiple value. Applying this method to our corpus and returning the best two hypothesis, we observe that the ground-truth tempo is present for 98 of the tracks. The global result of this method, choosing only the best comb as result, is 64% for $Accuracy_1$ and 96% for $Accuracy_2$. The detailled results are visible in Table 2.

Ratios with the correct tempo						
1/2	1	2	3	No link	Acc_1	Acc_2
3	64	29	0	4	64	96

Table 2. "Comb Decision" : Number of music tracks in function of the ratios between the estimated tempo and the ground truth value. $Accuracy_1$ and $Accuracy_2$ are deducted.

3.2.3 Combination of the strategies

As shown in Table 3, the combination of the two previous methods largely improves the results. The results in terms of $Accuracy_1$ is 78% and 93% in terms of $Accuracy_2$.

Ratios with the correct tempo						
1/2	1	2	3	No link	Acc_1	Acc_2
13	78	2	0	7	78	93

Table 3. Percentage of the returned values ratio of the ground truth for the Fusion of the two algorithms

The differences between their results is essentially due to the detection of the "double tempo". This type of error dissapears. The number of serious errors is stable.

3.3 Discussion

The 2004 MIREX evaluation was the last MIREX session which the task of tempo estimation was evaluated. These results were obtained on a corpus of 3199 tempo-annotated files ranging from 2 to 30 seconds, akd divided into three kinds : loops, ballroom music and songs exerpts.

The algorithms evaluated during this campaign are detailed and compared in [8]. The Klapuri's algorithm [9] obtained the best score on this evaluation with an $Accuracy_1$ of 67.29% and an $Accuracy_2$ of 85.01% among the total set of evaluated signals and reaching 91.18% of $Accuracy_2$ on the song's subset.

An exhaustive search for the best combination of five algorithms, using a voting mechanism, has also be computed. The best combination achieved 68% in terms of $Accuracy_1$, whereas the best $Accuracy_2$ reached 86%.

The MIREX corpus and the RWC part we use are different (in particular in terms of length). Nevertheless, our results are comparable and experiments will be realized on short extracts of the songs in order to define the robustness of our method.

4. CONCLUSIONS

In this paper, we presented a tempo estimator based on an automatic segmentation of the signal into quasi-stationnary zones. The use of this segmentation for the tempo induction seems to be rather significant: the spectrum of the Dirac signal derivate from the segmentation shows a predominant value directly linked with the tempo on 98% of our tests. The three methods which exploit this property have good performence. These methods are still rather simple, so we will investigate some potential improvements:

- Some experiments will be realized in order to evaluate the sensitiveness of our method to the use of short extract. Good results would allow the use of this method on slipping windows of few dozens of second. Such treatment could be realized in order to detect changes in the tempo.

- The use of the phase of $B_w(p)$ seems promissing for the developpement of a precise onset localizator.

5. REFERENCES

[1] M. Alonso, B. David, and G. Richard. Tempo and beat estimation of music signals. In *Proc. Int. Conf. Music Information Retrieval*, pages 158–163, 2004.

[2] R. André-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(1):29–40, 1988.

[3] J. Bilmes. Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, MIT, Cambridge, Mass., USA, 1993.

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

[5] S. Dixon, E. Pampalk, and G. Widmer. Classification of dance music by periodicity patterns. In *Proc. Int. Conf. Music Information Retrieval*, pages 159–165, 2003.

[6] M. Goto. Development of the RWC music database. In *Proceedings of the 18 th International Congress on Acoustics (ICA 2004)*, pages 553–556, 2004.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 287–288, 2002.

[8] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(5):1832 – 1844, september 2006.

[9] A. Klapuri, A. Eronen, and J. Astola. Analysis of the meter of accoustic musical signals. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(1):342–355, 2006.

[10] E. Scheirer. Tempo and beat analysis of acoustic music signals. *Journal of the Acoustical Society of America*, 104:588–601, 1998.

[11] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 10(5):293–302, 2002.

[12] C. Uhle, J. Rohden, M. Cremer, and J. Herre. Low complexity musical meter estimation from polyphonic music. In *Proc. AES 25th International Conference*, pages 63–68, June 2004.