TEMPO INDUCTION USING FILTERBANK ANALYSIS AND TONAL FEATURES

Aggelos Gkiokas[1,2], Vassilis Katsouros[1] and George Carayannis[2]
[1]Institute for Language and Speech Processing
[2]National Technical University of Athens
{agkiokas, vsk, gcara}@ilsp.gr

ABSTRACT

This paper presents an algorithm that extracts the tempo of a musical excerpt. The proposed system assumes a constant tempo and deals directly with the audio signal. A sliding window is applied to the signal and two feature classes are extracted. The first class is the log-energy of each band of a mel-scale triangular filterbank, a common feature vector used in various MIR applications. For the second class, a novel feature for the tempo induction task is presented; the strengths of the twelve western musical tones at all octaves are calculated for each audio frame, in a similar fashion with Pitch Class Profile. The time-evolving feature vectors are convolved with a bank of resonators, each resonator corresponding to a target tempo. Then the results of each feature class are combined to give the final output.

The algorithm was evaluated on the popular ISMIR 2004 Tempo Induction Evaluation Exchange Dataset. Results demonstrate that the superposition of the different types of features enhance the performance of the algorithm, which is in the current state-of-the-art algorithms of the tempo induction task.

1. INTRODUCTION

Tempo Induction has gained a great interest within the Music Information Retrieval community the past few years. Although in most systems, the tempo induction is made simultaneously with the beat tracking process as a unified task, the need for an individual handling of these tasks is apparent. An example can be found in Gouyon and Dixon in [1], where a genre classifier for 8 different music genres, based solely on the tempo of each excerpt has given remarkable results.

Beyond the scope of music classification, tempo induction and beat tracking are essential in many diverse applications, such as music similarity and recommendation, automatic transcription, audio editing, music to MIDI synchronization, and automatic accompaniment. They almost always serve as an inter-step in algorithms handling more complicated problems such as meter extraction [2] and rhythm description.

The algorithms that extract tempo can be divided into two main categories. The first consists of algorithms that use onset lists as input (either extracted directly from MIDI or audio). Indicative work can be found in [3], [4]. Most of these algorithms extract periodicities from the *inter-onset intervals* (IOIs) or by applying the autocorrelation function (ACF) to the onsets list in order to extract the tempo. In the latter belong the algorithms that search for periodicities directly from the audio (e.g. the ACF applied to frame features). Respective work can be found in [5], [6]. Although the former have the advantage of generalization (handling both MIDI and audio), evidence that the latter achieves better results is reported in [7]. An extensive review on the rhythm description algorithms can be found in [8].

A first step to systemize the tempo extraction task was the evaluation exchange organized during the 5th International Conference on Music Information Retrieval [7]. Seven participants submitted twelve different algorithms, tested on a collection of 3199 tempo-annotated music excerpts. The data was hidden from the participants. After the contest was conducted, the data were made available online (except of the *Loops* data that are available under a fee). Detailed description can be found in [7]. In a similar fashion, MIREX 2005[1] and MIREX 2006[2] Audio Tempo Extraction evaluation exchanges were conducted, with the difference that the evaluation procedure was more focused on the perceived than actual tempo. Unfortunately, the data is still not available except of a small portion that was used as training data.

Although a benchmark collection was created, few tempo induction algorithms have been tested on this dataset. A remarkable exception is Seyerlehner, Widmer and Schnitzer's work [9]. They proposed two versions of an algorithm that extracts rhythmic patterns using the autocorrelation function (ACF) as described in [10] and the Fluctuation Pattern as described in [11], respectively, in order to determine the tempo. Their approach is based on the assumption that pieces with similar rhythmic patterns are more likely to have similar tempo as well. The rhythmic patterns of excerpts are compared with those of a tempo-annotated music database. Their results showed that the proposed algorithm outperformed all the algorithms presented in the ISMIR 2004 evaluation exchange

[1] http://www.music-ir.org/mirex/2005/index.php/Audio_Tempo_Extraction
[2] http://www.music-ir.org/mirex/2006/index.php/Audio_Tempo_Extraction

on the *ballroom* data, and had similar results with the top performing algorithm [2] on the *songs excerpts* data. Note that the results presented are based on *accuracy1*, and not on *accuracy2*, where correct tempi are considered the fractions of the ground-truth tempo (half, double, three times, 1/3) which partly can be considered perceptually more relevant. Another example is the work of Alonso et al. in [12]. They proposed a system that estimates the tempo by decomposing the music signal sub-bands into harmonic and noise components. Then musical events are extracted with an "accentuation" weighting and periodicities are estimated. They tested the proposed algorithm on a corpus consisting of the *songs excerpts* collection and excerpts from the author's private collection. The evaluation measures were *accuracy1* and *accuracy2*, but with a 5% tolerance. Thus, their algorithm cannot be compared directly with the aforementioned.

In this paper we present a system that extracts tempo without onset detection, in a similar fashion that Scheirer does in [5]. The difference is that additionally to the filter-bank analysis, we incorporate a novel feature for the tempo induction task, similar to Pitch Class Profile, introduced by Fujishima [13]. The proposed algorithm assumes no significant tempo variation within the music excerpt.

The rest of the paper is organized as follows. In Section 2 we describe the architecture of our system. Section 3 focuses on details concerning the algorithm and individual processes of the implemented system. Comparable results on the *ballroom* and *songs excerpts* data of the ISMIR 2004 tempo induction evaluation exchange are provided in Section 4. Conclusions, drawbacks and future work conclude this paper in Section 5.

2. ALGORITHM OVERVIEW

The overall architecture and the individual components of the proposed algorithm are shown in Figure 1. Initially a moving Gaussian window is applied on to the input signal. For each frame a filterbank of equally spaced triangular filters in the mel-scale is applied, and the log-energy of each bank is calculated, in order to produce a vector **m** for each frame. Simultaneously, a similar process takes place, using a larger window. Each frame is convolved with twelve filters, each one corresponding to one of the twelve musical tones, forming the vector **t**.

Then a larger window of 8secs length is applied to each time-evolving feature, with a 1sec shift. Afterwards, the features are differentiated and convolved with a bank of resonators, with frequencies corresponding to the target tempi, and the $\|\bullet\|_\infty$ of each convolution is calculated. Then the norms are summed across features, independently for each feature type, forming two vectors $\mathbf{SC}_m$ and $\mathbf{SC}_t$ of length *T*, where *T* denotes the plurality of the target tempi. Each vector indicates the strength of each target tempo to the specific frame. Finally $\mathbf{SC}_m$ and $\mathbf{SC}_t$ are summed across the segments of the whole excerpt, to get the final tempo strengths for each feature class. The two vectors are combined to get the final output.

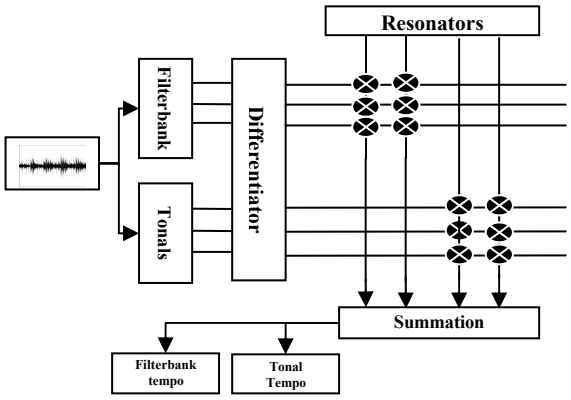

Figure 1. System Overview

3. ALGORITHM DETAILS

3.1 Extracting Filterbank Features

A moving Gaussian window of 20ms length with 5ms shift is applied to the input signal. Each segment is analyzed by a mel-scale triangular filterbank consisting of 12 bands, and the log-energy for each band is computed. This process forms a 12-dimensional feature vector **m**, for each frame.

3.2 Extracting Tonal Features

In a similar fashion with filterbank analysis, a Gaussian window is applied to the signal. In order to have better frequency resolution, the window is chosen much larger than the case before. A window with 80ms length and 5ms shift was chosen after experiments.

Each segment is convolved with 12 reference signals of same length with the sliding window. Each signal represents one the 12 western musical tones. The reference signals are the sum of cosines of equal amplitudes, at frequencies equal to the F_0 of each musical tone, at all octaves within a range from 27.5Hz to 10 kHz. No harmonic partials are considered. Formally, the reference signals are given by the following formula:

$$R_{tone(k)}(n) = \sum_{f_i \in \Omega_k} cos(2\pi f_i n), \quad k = 1..12 \qquad (1)$$

where Ω_k denotes the set of fundamental frequencies of tone *k* in the range of 27.5Hz to 10 kHz and *n* the time index. Afterwards the $\|\bullet\|_2$ of the twelve convolutions are calculated to form the tonal feature vector **t** for each frame. Formally

$$t_k(l) = \left\| (s(l) * R_{tone(k)})(n) \right\|_2, \quad k = 1...12 \quad (2)$$

where $s(n,l)$ is the signal frame l and $R_{tone(k)}$ the reference signal defined in Equation 1.

3.3 Convolving with the Bank of Resonators

The feature extraction process is followed by the convolution of the feature vectors with a bank of resonators. Firstly, we segment each time-evolving feature using a rectangular window of 8 secs with 1 sec overlap.

In order to compute the rhythmic periodicities of the signal, we convolve each feature segment with a bank of resonators, each resonator representing a specific tempo. We consider resonators with impulse response support that equals to the segment length for every integer value from 40 up to 280 bpm. The resonators can be any system with periodic impulse response, making our algorithm flexible to adapt to different signals. Motivated by the mathematical model of entrainment, as it was presented by Large and Kolen in [14], we adopted the basic oscillatory unit as the impulse response of the resonator. The equation of the oscillatory unit is given by

$$r(l) = 1 + \tanh(\gamma \cdot (\cos(2\pi\psi_t l) - 1)) \quad (3)$$

where ψ_t denotes the frequency of tempo *t*. Parameter γ is called the output gain.

After this process the $\|\bullet\|_\infty$ is computed for the convolution of every feature-resonator pair, resulting a vector indicating the strength of all tempi for each specific feature. Formally, we can write

$$S_{f_i}(t) = \left\| (r_t * f_i)(l) \right\|_\infty, \quad t \in T \quad (4)$$

where $S_{f_i}(t)$ is the strength of feature f_i at tempo *t* and T is the target tempi set. ($f_i \in \{\mathbf{m}_i, \mathbf{t}_i, i = 1..12\}$)

3.4 Combining the Feature Vectors

To compute the tempo for a specific segment, we summate the tempo strengths $S_{f_i}(t)$ across the features, *individually* for each feature class, thus taking two vectors $\mathbf{SC}_m$ and $\mathbf{SC}_t$, for filterbank and tonal features respectively. Formally, we can write

$$\mathbf{SC}_m(t) = \sum_{i=1}^{12} \mathbf{S}_{m_i}(t), \quad \mathbf{SC}_t(t) = \sum_{i=1}^{12} \mathbf{S}_{t_i}(t), \quad t \in T \quad (5)$$

where T is the target tempi set.

Finally, to combine the results of the two tempo detectors, we summate $\mathbf{SC}_m(t), \mathbf{SC}_t(t)$ across the segments of the music excerpt. Then by point-wise multiplication we compute the final vector **SC,** indicating the tempo strengths within the excerpt. The tempo with the maximum strength is considered as the correct tempo.

4. EXPERIMENTAL RESULTS

In this section we present the evaluation of the proposed algorithm. The data we used for the experiments consists of 1163 excerpts from the *ballroom* and *songs excerpts* datasets of the ISMIR 2004 Tempo Induction evaluation exchange. Details on the statistics, collection and annotation of the corpus can be found in [9].

Firstly, we evaluated our algorithm for each feature class individually. Afterwards we combined the outputs of the individual features as described in the previous section. The results on both *ballroom* and *songs excerpts* datasets are presented in Table 1.

	Ballroom		**Songs**	
Feature Type	**Acc1**	**Acc2**	**Acc1**	**Acc2**
Filterbank	56.34	93.33	23.01	88.39
Tonals	50.32	81.08	46.45	73.33
Combination	61.08	93.98	42.15	90.11

Table 1. Results (%) of the algorithm for the Ballroom and Songs Excerpts datasets, using feature classes individually and in combination .

It is clear that the algorithm yields better results using the filterbank features in the Ballroom dataset for the *accuracy1* measure. On the other hand, the algorithm performed poorly in the songs data based on *accuracy1* (only 23%). The above can be explained by the fact that the Ballroom data consists of more "percussive" excerpts, thus the filterbank energies represent sufficiently the data. Additionally, the experimental results demonstrate that by using solely the filterbank features, the proposed system "tends" to capture tempi double of the groundtruth tempo. For most of the excerpts classified correctly using *accuracy2* and misclassified using *accuracy1*, the detected tempo was double of the correct tempo.

When we used solely tonal features as input to the system, the *accuracy1* on the songs data increased significantly (from 23% to 46.5%) for the *Songs* data. This can be explained by the more "melodic" nature of the excerpts consisting *Songs* data, which prove tonal features to be more suitable for that case. On the other hand *accuracy2* measure degraded from 88.39% to 73.3%. A possible explanation is the large window used in the preprocessing stage, which "cuts off" frequencies double or triple of the actual tempo.

When combining the results from the two versions, we observe that in both Ballroom and Songs excerpts data, the superposition increases the algorithm performance, especially in the Songs Data. Considering the filterbank features as base features, tonal features provide additional information about the rhythm periodicities of the signal. Comparative results of the presented algorithm, namely GK, with the best five performing algorithms in [7], namely *Miguel Alonso* (AL), *Simon Dixon* (DI), *Anssi*

Klapuri (KL), *Christian Uhle* (UH) and *Eric Scheirer* (SC), plus *Klaus Seyerlehner* (SE1,SE2)[9] are presented in Table 2.

	Ballroom		Songs	
Method	**Acc1**	**Acc2**	**Acc1**	**Acc2**
GK	61.08	93.98	42.15	90.11
AL	34.1	69.48	37.42	68.6
DI	43.12	86.96	16.99	76.99
KL	63.18	90.97	58.49	91.18
UH	56.45	81.09	41.94	71.83
SC	51.86	75.07	37.85	69.46
SE1	78.51	-	40.86	-
SE2	73.78	-	60.43	-

Table 2. Comparative results (%) on *Ballroom* and *Songs* datasets.

5. CONCLUSION AND FURTHER WORK

In this paper we presented a system that extracts the tempo of a music signal. The proposed algorithm was evaluated on the benchmark corpus of the ISMIR 2004 with encouraging results. Without taking into consideration any high-level musical information, our system performed within the current state-of-the-art algorithms of the tempo induction task.

The tonal features introduced in this work prove to capture additional aspects of rhythmic periodicity in a musical signal. It is evident that underlying rhythmic periodicities of a musical signal can be found beyond the filterbank energies, in a more "pitched context". Without any multi-pitch estimation or chord detection process, we observe that the simpler and more abstract tonal features presented in this paper similar to Pitch Class Profile, contain rhythmic information that can enhance the performance of a tempo induction system that does not take into account any tonal information.

However, during the experiments we observed that the performance of the presented algorithm is sensitive to the window length and shift during the extraction process of tonal features, an effect that will be investigated in the future. Moreover we intend to extend tonal features in a more sophisticated way, such as chords, and incorporate harmonic partials information. Finally, the superposition of the output for the features classes is a subject for future research.

6. REFERENCES

[1] Gouyon F. and Dixon S., "Dance Music Classification: A Tempo-Based Approach", *Proceedings of the 5th International Conference on Music Information Retrieval,* Barcelona, Spain 2004

[2] Klapuri A., Eronen A. and Astola J., "Analysis of the Meter of Music Acoustic Signals", *IEEE Trans. Audio, Speech, and Language Processing, 14(1),* January 2006.

[3] Alonso M., David B., Richerd G., "Tempo and Beat Estimation of Musical Signals", *Proceedings of the 5th International Conference on Music Information Retrieval,* Barcelona, Spain 2004.

[4] Davies M., Plumbley M., "Context-Dependent Beat Tracking of Musical Audio", *IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 3*, March 2007.

[5] Scheirer E., "Tempo and Beat Analysis of Acoustic Musical Signals.", *The Journal of the Acoustical Society of America, Vol. 103, No. 1,* January 1998.

[6] Dannenberg R., "Toward Automated Holistic Beat Tracking, Music Analysis, and Understanding", *Proceedings of the 6th International Conference on Music Information Retrieval, ,* London, UK, 2005.

[7] Gouyon F., Klapuri A., Dixon S., Alonso M., Tzanetakis G., Uhle C., and Cano P., "An Experimental Comparison of Audio Tempo Induction Algorithms", *IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 5,* September 2006.

[8] Gouyon F. and Dixon S., "A Review of Automatic Rhythm Description Systems", *Computer Music Journal, 29:1, pp 34-54,* Spring 2005.

[9] Seyerlehner K., Widmer G., and Schnitzer D., "From Rhythm Patterns to Perceived Tempo", *Proceedings of the 8th International Conference on Music Information Retrieval, ,* Vienna, Austria, 2007.

[10] Ellis D.P.W. "Beat Tracking with Dynamic Programming", *Journal of New Music Research,* vol. 36 no. 1, March 2007, pp 51-60.

[11] Pampalk E., Rauber A., Merkl D., "Content-Based Organization and Visualization of Music Archives", *Proceedings of the 10th ACM International Conference on Multimedia*, Juan les Pins, France, 2002.

[12] Alonso M., Richard G., David B., "Accurate Tempo Estimation Based on Harmonic + Noise Decomposition", *EURASIP Journal on Applied Signal Processing Volume 2007, Issue 1,* January 2007.

[13] Fujishima T. "realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music".

[14] Large E. and Kolen J., "Resonance and the Perception of Musical Meter", *Connection Science 6(1), pp 177-208,* 1994.