TEMPO ESTIMATION BASED ON LINEAR PREDICTION AND PERCEPTUAL MODELLING

Georgina Tryfou[1,2], **Aki Härmä**[3], **Athanasios Mouchtaris**[1,2]

[1]Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH-ICS), Heraklion, Crete, Greece
[2]Department of Computer Science, University of Crete, Heraklion, Crete, Greece
[3] Philips Research, Eindhoven, The Netherlands
tryfou@ics.forth.gr, aki.harma@philips.com, mouchtar@ics.forth.gr

ABSTRACT

Many applications demand the automatic induction of the tempo of a musical excerpt. The tempo estimation systems follow a general scheme that consists of two main steps: the creation of a feature list and the detection of periodicities on this list. In this study, we propose a new method for the implementation of the first step, along with the addition of a final step that will enhance the tempo estimation procedure. The proposed method for the extraction of the feature list is based on Gammatone subspace analysis and Linear Prediction Error Filters (LPEFs). As a final step on the system, the application of a model that approximates the tempo perception by human listeners is proposed. The results of the evaluation indicate the proposed method compares favourably with other, state-of-the-art tempo estimation methods, using only one frame of the musical experts when most of the literature methods demand the processing of the whole piece.

1. INTRODUCTION

The tempo is a dominant element connected to the hierarchical structure of a music signal that can define various aspects of it. Moreover, it is an intuitive music property that human listeners, even without any musical education are able to perceive and understand only by listening to the first few seconds of an excerpt. The tempo is defined as the rate of the *tactus* pulse, a prominent level in the hierarchical structure of music, which is also referred to as the *foot-tapping* rate.

The process of automatically inferring the tempo of a musical piece plays an important role among the applications in the field of Music Information Retrieval (MIR). Many of them, for example, beat tracking and music classification, need a preprocessing stage where tempo estimation takes place. Beyond these, tempo induction is essential in music similarity and recommendation, automatic transcription and even audio editing. More complicated tasks such as meter extraction and rhythm description also demand a tempo estimation module. Finally, in applications with beat synchronous visual and audio effects the estimation of the tempo is a necessary part.

In such applications it is desired that correct tempo estimation would be available to the system at about the same time that the tempo is detected by a human listener. This is technically very difficult because the human listeners are able to use higher-level context cues to conduct tempo detection. In fact, many algorithms proposed for tempo estimation in the past [7, 11] require a long signal segment for producing reliable results. This is clearly a problem in contents such as radio programs, where the rhythmic music content may alternate with, for example, speech segments

Tempo induction algorithms follow a general scheme [4, 5], that consists of two main stages. In the first stage, the audio signal is parsed and a set of features is created. These features convey an initial rhythmic structure of the input musical piece. Literature reveals two main methods to obtain features: either from a list of the *inter onset intervals* (IOIs) of the musical signal or from the temporal evolution of the musical signal.

Representative algorithms that fall in the first category, and use IOIs for the creation of the feature list, are presented in [1–3]. Algorithms in the second category rely on features extracted directly from the audio signal. These features may emphasize onset locations but they do not result from onset lists. In [11] an amplitude envelope of the signal in six octave-spaced subbands is created at the first stage of the system. This approach is expanded in [7], where a more

generic and therefore robust *accent signal* is created across four subbands.

During the second stage of the tempo estimation scheme, periodic recurrences of the features are found and the tempo is calculated. There are several methods to achieve this. For example, the autocorrelation function (ACF) [12], comb-filter resonators [7, 11] and phase-locking resonators [8].

The method proposed in this paper uses Gammatone analysis and linear prediction for extracting the feature list. After that, the estimation of the existing periodicities takes place. As a last step to the algorithm, a perceptual weighting method is proposed for enhancing the system's accuracy.

The results of the system are encouraging and indicate that the addition of a perceptually inspired stage at the end is advantageous for an algorithm that follows the tempo estimation general architecture. In addition to that, the use of a single, 4 seconds long, frame to obtain the final results, facilitates the proposed algorithm to quickly adapt to possible tempo changes in a given music excerpt.

The Gammatone filterbank models the input signal using a frequency resolution which is similar to that of the human auditory system. Moreover, the use of LPEFs in the first step, enables the accentuation of points where abrupt changes take place in any frequency band of the input signal. These points in time are considered significant for the task of tempo estimation.

The perceptual processing, that starts with the application of the Gammatone filterbank on the input signal, proceeds with the weighting method that is added as a last step on the system. This weighting method is based on a resonance model that has been found to follow the perceptual responses to a variety of musical excerpts [10, 13]. The use of this model in order to enhance a tempo estimation system is novel and leads to promising results.

The rest of the paper is organised as follows. In Section 2 the architecture of the system is described. Section 3 provides results and evaluates the developed system. Finally, conclusions and future work are discussed in Section 4.

2. METHOD DESCRIPTION

The developed system follows the general scheme of tempo estimation algorithms, with the addition of a last step where the perceptual processing of the results takes place. The block diagram of the system is shown in Figure 1. Each one of the three major units depicted there, is described in details in the following sections.

2.1 Feature List Extraction

When a listener listens to music, the musical events are related to a regular pattern of beats, called *metrical structure*. These patterns are organised in a metrical hierarchy that exists in every musical sound and consists of two or more levels. When a beat is felt stronger that the other beats of the same metrical level then it is also a beat at the higher musical level. This hierarchy is depicted in Figure 2, where also the first three levels of it are presented. The tempo is described as the rate of the *tactus* beat, or based on the above explanation the rate at which strong beats appear at the *tatum* level.

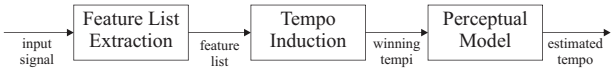

Figure 1. The block diagram of the proposed system.

Figure 2. The hierarchical structure of a piece with a 4/4 meter

During this stage of the analysis the goal is to detect events that are connected to the strong beat of the *tatum* level. To achieve this, it is assumed that any event perceived as a strong beat will appear as an abrupt increase in the temporal evolution of the musical signal, baring significantly more transient content than the rest of the beat-related events.

Let us consider the input music signal $x[n]$. The first step of the processing is the application of a bank of K Gammatone filters on it:

$$x_k[n] = h_k[n] \star x[n] \quad k \in [0, 1, \ldots, K-1], \qquad (1)$$

where $h_k[n]$ the impulse response of the k-th Gammatone filter. During the implementation the value K was chosen to be 16.

After the filtering of the input signal, each subband signal is decimated K times as follows

$$x_k[n] = x_k[Kn]. \qquad (2)$$

The $x_k[n]$ signals are then given as an input to a bank of adaptive LPEFs. The use of the LPEFs enables the detection of abrupt changes in the temporal evolution of the signal. By adapting the linear prediction coefficients that these filters use, it is possible to emphasize the events that the adaptive algorithm fails to model. The strong beats that appear at the *tatum* level are connected to these events.

The output of the adaptive LPEFs is the prediction error of the adaptive linear predictive algorithm given in Algorithm 1. This algorithm is based on estimating the LPC

coefficients of the initial M values of the N long frame, and adapting these coefficients using the Least Mean Squares (LMS) algorithm for the remaining $N - M$ samples. The selected values for M and N are 23 ms and 1 second respectively (converted in samples). More details on linear prediction and the LMS algorithm can be found in [6].

The output signal, $df_k[n]$, is the *detection function*, a residual signal that presents high values when beat related events take place in the temporal evolution of the signal.

Algorithm 1 The implemented adaptive LPEF algorithm.

$mu \leftarrow 10^{-3}$
$\mathbf{w}_k[0] \leftarrow \text{LPC}(\mathbf{x}_k[0])$
for $n = 1$ **to** $N - M$ **do**
  $\tilde{x}_k[n] \leftarrow \mathbf{w}_k^T[n] \star \mathbf{x}_k[n]$
  $df_k[n] \leftarrow x_k[n] - \tilde{x}_k[n]$
  $\mu \leftarrow \min\left(mu, \frac{1}{x_k^T[n]x_k[n]}\right)$
  $\mathbf{w}_k[n+1] \leftarrow \mathbf{w}_k[n] + \mu df_k[n]x_k[n]$
  $n \leftarrow n + 1$
end for

A peak picking procedure, applied on the analysis frames (of length N) of the smoothed and normalized signals $df_k[n]$, produces a time series

$$ts_k[n] = \begin{cases} 1 & \text{if } df_k[n] \text{ demonstrates a peak here} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

During peak picking, an adaptive threshold, calculated by the sum of a predefined, static threshold and a moving median filter is used.

The time series $ts_k[n]$ are then convolved with a Hanning window in order to produce the *mask functions* $m_k[n]$. In that way, a strongly smoothed version of the corresponding *detection function* is created that however accentuates the detected abrupt events. The above described processing for the creation of the feature list combines the advantages of the use of an onset list with those methods where the feature lists are obtained in a continuous manner.

2.2 Tempo Induction

In the second part of the system, the periodicity analysis is carried out, in order to infer the tempo from the list of features (*i.e.* mask functions). The periodicity analysis is done using a bank of comb filters.

Each one of the *mask functions*, $m_k[n]$ is given as an input to a bank of comb filters. Therefore, for the analysis band k the following takes place:

$$y_{k,\tau}[n] = a_\tau y_{k,\tau}[n - \tau] + (1 - a_\tau)m_k[n], \quad (4)$$

for every $\tau \in \mathcal{T}$. The interval $\mathcal{T}$ ranges from 42 to 242 *beats per minute* (BMP). In this interval the filter's delay τ takes integer values. The term a_τ corresponds to the filter's feedback gain and it is calculated as $a = 0.5^{\frac{\tau}{T_0}}$. The time during which the signal should reach its half energy is T_0. In this system T_0 is equal to 4 seconds. The selection of this time frame is motivated by the smallest tempi normally found in a piece of music. With a minimum tempo of 42 BPM, this frame is big enough to cover at least two repetitions of the beat but also small enough for the system to quickly adapt to any tempo changes, when more than one frames are used as an input.

The energy of each filter, in each frequency band k is then calculated by

$$e_{k,\tau}[n] = \frac{1}{\tau} \sum_{i=n-\tau+1}^{n} y_{k,\tau}[i]^2. \quad (5)$$

A sum across all the frequency bands k will result to a wide band energy signal for each tempo τ

$$e_\tau[n] = \sum_{k=1}^{K} e_{k,\tau}[n] \quad (6)$$

So far, for every time index n of the input signal we obtain a vector

$$\mathbf{e} = [e_{42}[n] \quad e_{43}[n] \quad \dots \quad e_{242}[n]]^T \quad (7)$$

consisting of the instant energies in every periodicity $\tau \in \mathcal{T}$.

The N_T maximum components of the vector $\mathbf{e}$ are then selected in order to form a vector $\mathbf{w}$. The corresponding tempi form the vector $\mathbf{T}$. The vector $\mathbf{T}$ contains the *winning tempi*, and vector $\mathbf{w}$ their relative weights.

2.3 Perceptual Model

The ambiguity in the perception of tempo has been modelled and tested in experiments [10, 13] where the distribution of responses from several listeners to the same pieces of music were studied. This analysis resulted in the following resonance model:

$$A_e(t) = \frac{1}{\sqrt{(t_o^2 - t^2)^2 + \beta t^2}} - \frac{1}{\sqrt{t_o^4 + t^4}}, \quad (8)$$

where A_e is the effective resonance amplitude, t_o is the resonance tempo, β the damping constant and t is the tempo variable. During experimenting, these parameters were fitted to the distribution of the tapped tempi. It has been found that, on average, music experts produce a resonant tempo of 138 BPM with a damping constant, β equal to 5.0. In Figure 3 the produced model, with the use of these parameters is depicted.

In this paper the model of equation (8) is used to weight the results from the periodicity analysis so that

$$w_i' = A_e(T_i)w_i \quad i \in [1, 2, \dots, N_T], \quad (9)$$

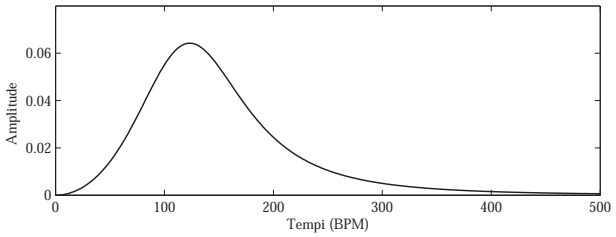

Figure 3. The resonance model that was described in [10] and fits the distributions of responses to several pieces of music.

where T_i the i-th value of vector $\mathbf{T}$ and w_i the corresponding weight. After this step, elements w'_i form $\mathbf{w}'$ which contains the perceptually modified weights of the winning tempi in $\mathbf{T}$. The tempo estimation is therefore enhanced with perceptual information.

Systems that estimate periodicity patterns in a signal strongly respond to the multiples and aliquots of any fundamental periodicity that appears in it. Likewise, when it comes to human listeners, the more ambiguities in determining the tempo appear due to the selection of multiples and divisors of the same tempo. In the vector of winning tempi, $\mathbf{T}$, also appear not only possible perceived tempi, but also multiples and aliquots of them.

In order to discard some "false" estimations from $\mathbf{T}$ and decide which is the perceived tempo in a group of tempi that have a common divisor, an extra weighting step is introduced. During experimenting, it was found that increasing the weights of each tempo that appears in $\mathbf{T}$ with a factor of the weight of its multiples and divisors that also appear in $\mathbf{T}$, has the following two desired effects:

a. Significant decrease in the (normalized) weights of tempi whose multiples and aliquots are not present.

b. Highly accurate decision on which is the true perceived tempo within a set of tempi that have the same common divisor (as presented in Section 3).

This factor was chosen experimentally 0.3 for multiple periods and 0.6 for aliquots.

3. EVALUATION AND RESULTS

3.1 Datasets and Evaluation Measures

The developed system is evaluated using the measures proposed in [5]. The two measures defined are *Accuracy 1* and *Accuracy 2*, corresponding to the percentage of tempo estimates within 4% of the ground truth data. For the calculation of *Accuracy 2* also integer multiplications and divisions of the ground-truth tempo are considered to be correct estimates.

Winning Tempi	**T1**	**T2**	**T3**	**T4**	**T5**
Accuracy 1 (%)	38.40	28.22	6.02	1.72	2.44
Accuracy 1 CDF (%)	38.40	66.62	72.64	74.36	76.80

Table 1. The *Accuracy 1* of the algorithm for the estimation of the winning tempi

The results are based in two different datasets, both used in [5] for a comparative evaluation of tempo induction algorithms. That way our results can be compared to previous work. The first dataset, *Ballroom*, consists of 698, (30 seconds long each) audio excerpts. The second dataset, *songs*, contains 465 audio excerpts, this time each one being around 20 seconds long. The two datasets cover a wide range of genres (namely Rock, Classic, Electronica, Flamenco, Jazz, AfroBeat, Samba, Balkan, Greek, Cha Cha, Rumba, Samba, Jive, Quickstep, Tango and Waltz). Both datasets have been made publicly available [1] . It is mentioned here that due to some missing or bad formatted files, the following results have been calculated over a subset of the above datasets, that covers the 97.25% of the whole data.

3.2 Results

The first phase of the evaluation procedure was to check the accuracy of the algorithm in defining the vector of the winning tempi, $\mathbf{T}$, *i.e.* before applying the perceptual modelling. The winning tempi in the vector are placed in descending order, based on their weight. In Table 1, the results on the *Ballroom* dataset are illustrated. In the first row, the *Accuracy 1* of the algorithm in each index of the winning tempi is shown. The next row, presents the cumulative results up to each index of the vector $\mathbf{T}$. As depicted in this table, the algorithm has a success rate of 76.8% in estimating the correct tempo in the first 5 estimations.

The perceptual model at the end was inspired by this ambiguity in the results. Although the algorithm is quite accurate in detecting the right periodicity from a music excerpt, it has a relatively low percentage (38.4%) to do so in the first guess (*i.e.* the tempo with the higher energy). Until this point, only low-level music features have been used. The encoding of higher level knowledge on tempo perception in the model could be useful in choosing the right index of the winning tempi vector as the final estimation.

Indeed, the last step of the system achieves this task. The perceptual method is applied to the output, improving significantly the results of the algorithm. The results on both datasets, for the two evaluation metrics can be seen in Table

[1] http://mtg.upf.edu/ismir2004/contest/tempoContest/

	Ballroom		Songs	
Method	A1	A2	A1	A2
Simple	38.40	69.05	34.68	55.18
Perceptual	57.31	80.80	51.80	69.14

Table 2. Resulting percentages of the algorithm

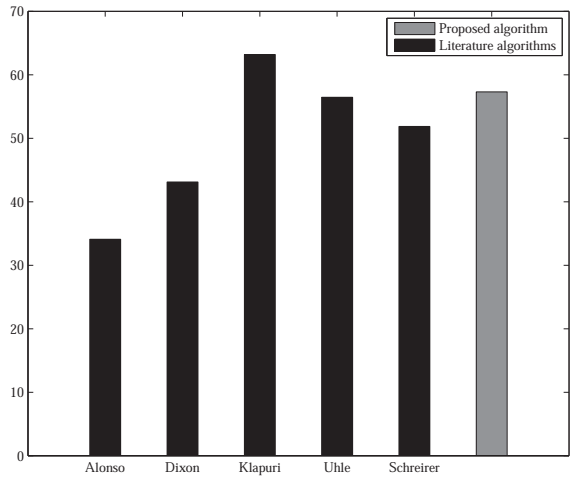

Figure 4. *Accuracy 1* on the *Ballroom* dataset. The literature algorithms mentioned are the following: Alonso [1], Dixon [3], Klapuri [7], Uhle [12], and Scheirer [11].

2. In the first row the results of the two datasets are presented, for both measures *Accuracy 1* (A1) and *Accuracy 2* (A2), without the application of the perceptual modelling described in Section 2.3. In the second row, the corresponding accuracy values are shown after the application of the perceptual weighting.

Comparing Table 1 and Table 2, it becomes clear that there is a significant improvement of 49% in the *Accuracy 1* measure when the perceptual weighting is used as a last step. Moreover, the fact that this improvement is not followed in *Accuracy 2* measure implies that the improvement in estimation takes places due to less multiplication and division errors.

As mentioned above, the use of the *Ballroom* and *Songs* datasets, along with the use of the *Accuracy 1* and *Accuracy 2* measures, enables the comparison of the results to the current state-of-the-art algorithms. In Figure 4 such a comparison is depicted and the proposed system seems to perform well. Further improvements to the proposed method are envisioned and these are discussed in the following section.

4. CONCLUSIONS AND FUTURE WORK

A new method to estimate the tempo of musical signals was presented in this paper. The evaluation of this method was conducted using popular datasets for the tempo estimation task along with previously defined evaluation measures. Although at an early stage, the algorithm seems to operate very well in comparison to the state-of-the-art, using only a single frame (4 seconds long) for calculating the result.

As mentioned, the above described results are obtained from a single frame of the input signal. An application of the algorithm on the whole signal, and then the computation of a median or average tempo estimate did not seem to yield a significant improvement. However, the implementation of a voting mechanism could improve the overall tempo estimate of a piece. In such an extension an extra assumption has to be made, *i.e.* that the tempo of the piece does not present any variations throughout the song.

The use of adaptive LPEFs introduced by this work, seems to work well in the task of extracting tempo estimation features. However, it was observed during experimenting, that the final results and success rates are sensitive to the the set of parameters used by the feature list extraction part (LPC order, peak picking static threshold). A detailed examination of the results that are obtained from different parameter sets and the determination of an optimum set may further improve the accuracy of the whole system.

Furthermore, the use of a different set of temporal features that indicate the tempo can be considered in a later version of the algorithm as the literature reveals some promising alternatives. For example, linear prediction coefficients instead of the the prediction error have been successfully used as features for music genre classification in [9].

Until now, the existing knowledge on the perceptual event that leads to the well known action of *foot-tapping*, has not been extensively used for a systematic way of estimating perceived tempo. This study indicates that taking advantage of auditory modelling tools can significantly improve the performance of a tempo estimation algorithm.

5. ACKNOWLEDGEMENT

This work has been funded in part by the European Community's Seventh Framework Programme under grant agreement n^o 230709 (PEOPLE-IAPP "AVID-MODE" grant).

6. REFERENCES

[1] M. Alonso, B. David, and G. Richard. Tempo and beat estimation of musical signals. In *Proceedings of the 5th International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[2] M. P. E. Davies and M. D. Plumbley. Context-dependent beat tracking of musical audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1009–1020, 2007.

[3] S. Dixon, E. Pampalk, and Widmer G. Classification of dance music by periodicity patterns. In *Proceedings of the International Conference on Music Information Retrieval*, pages 159–165, 2003.

[4] F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54, 2005.

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

[6] S. Haykin. *Adaptive filter theory*. Prentice Hall, 1995.

[7] A. P. Klapuri, A. J. Eronen, and J. T. Astola. Analysis of the meter of acoustic musical signals. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(1), 2006.

[8] E. W. Large and J. F. Kolen. Resonance and the perception of musical meter. *Connection Science*, 6(1):177–208, 1994.

[9] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen. Temporal feature integration for music genre classification. *Audio, Speech and Language Processing, IEEE Transactions on*, 15(5):1654–1663, 2007.

[10] D. Moelants and M. F. McKinney. Tempo perception and musical content: What makes a piece fast, slow or temporally ambiguous. In *Proceedings of the 8th International Conference on Music Perception and Cognition*, 2004.

[11] E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *Acoustical Society of America*, 103, 1998.

[12] C. Uhle, J. Rohden, M. Cremer, and J. Herre. Low complexity musical meter estimation from polyphonic music. In *Audio Engineering Society Conference: 25th International Conference: Metadata for Audio*, pages 63–68, New York, 2004.

[13] L. Van Noorder and D. Moelants. Resonance and the perception of musical meter. *Journal of New Music Research*, 28(1):43–66, 1999.