HAL open science

Musical tempo estimation using noise subspace projections

Miguel Alonso Arevalo, Roland Badeau, Bertrand David, Gael Richard

To cite this version:

Miguel Alonso Arevalo, Roland Badeau, Bertrand David, Gael Richard. Musical tempo estimation using noise subspace projections. Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2003, New Paltz, New York, United States. pp.95–98. hal-00945267

HAL Id: hal-00945267

https://inria.hal.science/hal-00945267v1

Submitted on 24 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MUSICAL TEMPO ESTIMATION USING NOISE SUBSPACE PROJECTIONS

Miguel Alonso, Roland Badeau, Bertrand David and Gaël Richard

ENST, Département de Traitement du Signal et des Images
46, rue Barrault, 75634 Paris cedex 13, France
{malonso,rbadeau,bedavid,grichard}@tsi.enst.fr

ABSTRACT

Tempo estimation plays a fundamental role in music analysis, especially for the automatic processing of large amounts of musical data. In this paper, a novel idea to enhance the estimation of the tempo in musical pieces is described, based on an harmonic/noise decomposition. This separation of the signal into a deterministic and a stochastic part is performed by projecting the signal onto its noise subspace. Besides, the proposed algorithm shares various elements with other tempo estimation methods. On a database composed of 54 excerpts from many musical genres our algorithm scored a success rate of 96%.

1. INTRODUCTION

Despite the apparent simplicity of the foot-tapping or beat-tracking skill compared to other musical ones, computer-based tempo estimation has remained a difficult work for processing a wide range of musical genres. The tempo, also known as the *beat*, is an important rhythmic property of music. In this paper, an algorithm is described, that determines the period between two musical motifs or beats and finds the rate in beats per minute (BPM) at which they occur. There exist numerous examples of applications: musical analysis, automatic rhythm alignment of multiple musical instruments, audio content analysis for automatic indexing and retrieval, beat driven special effects, musical education, etc. So far, some of the tempo tracking systems found in the literature focus on finding the tempo of strong beat[1] signals [1–3]. Some are more general, but are often challenged by orchestral classical music, mainly because of the weakness of the attacks and tempo variations [4, 5].

Many tempo estimation algorithms share the same principle. First, they decompose the input signal into a number of frequency bands by means of a filter bank [1, 4, 5] or by grouping frequency bins in the DFT of the signal [2, 3]. No consensus has been established about an optimal frequency decomposition and according to experimental results presented in [1], many frequency decompositions lead to satisfactory results. Nevertheless, there is a tendency towards splitting the signal in a logarithmic scale [1, 4, 5]. Next, extraction of subband onsets is carried out. As for the previous stage, there are different approaches to solve this problem. In [2, 3] onset detection is performed independently in clusters of frequency bins directly from the spectrogram, in the first case using the first-order difference function and in the second case using a tailored-made function. In [1, 4, 5] onsets are detected by computing the bandwise temporal envelope (the signal is rectified, lowpass filtered and decimated) and then calculating the first-order difference function as in [1], or a slight variation of the relative first-order difference function [5] introduced in [6]. Although the envelope extraction is rather similar in most systems, there is a large variety of methods employed to estimate subband periodicity. For example, in [1] the author uses a bank of oscillators which resonate at integer multiples of their characteristic frequency, [4] uses a fundamental frequency estimation method (called YIN [7]) which is a more robust replacement of the autocorrelation function, [2] uses a four component Gaussian mixture model to express the likelihood of onset locations, others compute in each frequency band an inter onset interval (IOI) histogram to determine the tempo [3] or the *tatum*[2] [5].

In this paper, we assume that the tempo of the audio signal is constant over the duration of the analysis window. The proposed system aims at estimating the tempo for a variety of musical genres. Its performance was tested using an experimental manually annotated data base comprising excerpts from rock, latin (cuban/salsa), pop music, jazz, classical and traditional songs. The paper is organized as follows: section 2 provides a detailed description of the proposed system, except for the noise subspace projection algorithm which is described in section 3. Next, in section 4 tests results are provided and compared to other existing methods, issues about the sensitivity and robustness of the presented method are also addressed. Finally, section 5 summarizes the main achievements of the presented algorithm and gives some possible directions for future research.

[1] According to [1], regardless of their melodic complexity strong beat signals bear a straightforward, perceptually simple rhythm.

[2] The term "tatum" has been derived from *temporal atom* and is the smallest time interval between successive notes in a rhythmic phrase.

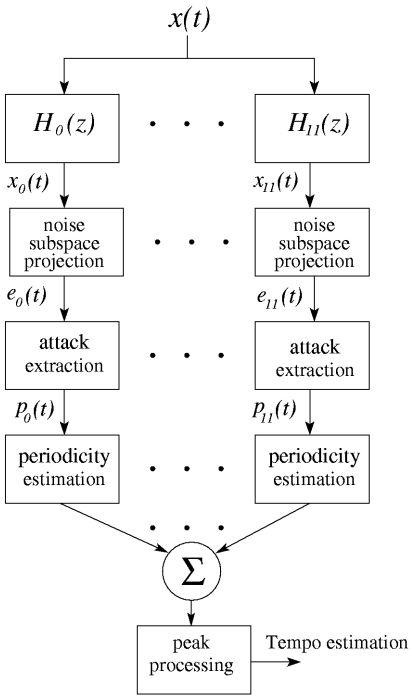

Figure 1: *Overview of the system.*

2. DESCRIPTION OF THE ALGORITHM

Our tempo estimation algorithm shares a number of features with other systems found in the literature. The main difference is that we estimate the tempo from the residual component of a harmonic plus noise model, in order to highlight the attacks in the musical signal.

The General overview of the tempo estimation system is presented in Figure 1. The algorithm works as follows: the input audio signal $x(t)$ is first decomposed into twelve uniform non-overlapping subbands using a cosine modulated filterbank where the prototype filter is implemented using a 200th order FIR filter with 80 dB of rejection in the stop band. The use of a highly selective filter is necessary because the noise subspace projection stage is very sensitive to spurious sinusoids in the stop band. The filterbank output signals $x_k(t)$, where $k = 0, ..., 11$, are then projected onto the noise subspace (*cf.* section 3), providing the corresponding residuals $e_k(t)$.

The subband onset detection is the next stage, once the bandwise noise signals $e_k(t)$ have been calculated. A compound system is used, which combines the schemes proposed in [6, 8]. Figure 2 depicts the flow diagram of this onset detector. The role of the filter $H_k(z)$ is described as follows. After subtracting the harmonic components from the subband signal $x_k(t)$ the energy ratio between the pass band and the stop band for the noise signals $e_k(t)$ has been severely reduced. Therefore, prior to further processing it is necessary to refilter those signals in order to eliminate noise components outside the desired frequency range. The next three diagram blocks concern the envelope extraction. To compute the subband noise envelope $a_k(t)$, the noise signal $e_k(t)$ is half-wave rectified and low-pass filtered using a 100ms decreasing waveform defined as the 2nd half of a 200ms Hanning window. The narrow-band output signal is then decimated by a factor of $M = 16$ for the purpose of reducing the computational burden. To find the onset points in $a_k(t)$, the first order relative difference function $w_k(t)$ proposed in [6] is used. This function gives the amount of change in the signal in relation to its absolute level. This is equivalent to differentiating the logarithm of the subband noise envelope, as given by Eq. (1).

$$w_k(t) = \frac{d}{dt} \ln(a_k(t)) \tag{1}$$

The first order relative difference function is a psychoacoustically relevant measure, since the perceived increase in signal amplitude is in relation to its level, the same amount of increase being more prominent in signals with a low dynamical range [6]. Hence, we find onset points by a peak picking operation, which looks for peaks above a given threshold. An adequate threshold value was found experimentally to be near $1.5\,\sigma_{w_k}$, where σ_{w_k} stands for the standard deviation of $w_k(t)$.

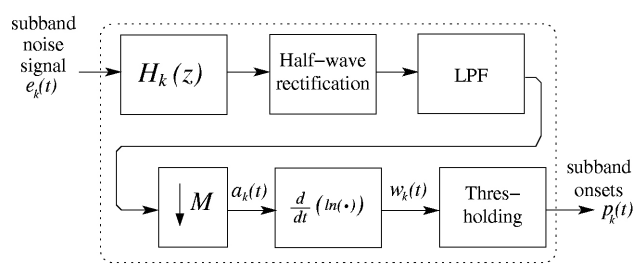

Figure 2: *Attack extraction flow diagram.*

Periodicity estimation. The signal $p_k(t)$ at the output of the onset detection stage is a train pulse. It can be seen as composed of two parts: one corresponding to a quasi-periodic pulse signal, bearing the *beat* of $x(t)$, and an additive noisy pulse signal. To ease the periodicity estimation of the train pulse, the signal $p_k(t)$ is convolved with a 75ms odd-length Hanning window. This leads to a smooth quasi-periodic signal $z_k(t)$. The method we employed to determine the periodicity of $z_k(t)$ was motivated by the work presented in [8], which is a very efficient model in time-domain pitch analysis, and is based on the summary autocorrelation function (SACF). First, the signal $z_k(t)$ is centered ($\tilde{z}_k(t) = z_k(t) - \bar{z}_k$) and then the autocorrelation function (ACF) is calculated, as shown in Eq. (2)

$$r_k(\tau) = \sum_t \tilde{z}_k(t+\tau)\tilde{z}_k(t) \quad (2)$$

where $0 \leq t \leq W-1$, and W stands for the analysis window length. Finally, the ACFs r_k are summed across bands to form the summary ACF (SACF).

We suppose that the tempo lies between 60 and 200 BPM, without loss of generality since any other value can be mapped into this range. Hence, during the SACF calculation, it is merely necessary to shift τ within the 300 ms to 1,000 ms time range. To find the periodicity of the subband envelope signals $z_k(\tau)$, we look for the three most salient peaks in the SACF. To ensure a more robust and accurate tempo extraction, a multiplicity relationship between them is extensively searched via a numerical algorithm. If no multiplicity relation is found among the detected peaks, the time lag of the most prominent one is taken as the beat period.

3. NOISE SUBSPACE PROJECTION

The noise subspace projection stage is processed independently in each subband. It is based on the *Exponentially Damped Sinusoidal* (EDS) model [9]:

- The *harmonic* part of the sound is modeled as a sum of n sinusoids, which may have an exponential decay. In order to include polyphonic sounds, the frequencies of these sinusoids are not constrained to be evenly distributed.
- The *noise* part is defined as the difference between the original signal and the harmonic part.

The estimation of multiple sinusoids in noise has been extensively investigated in signal processing literature [10]. Among the various approaches, subspace-based techniques are of major interest because they overcome the resolution limit of the Fourier analysis, they are robust to high noise levels, and they can be used with short analysis windows.

Consecutive snapshot vectors of length L are extracted from the signal. The subspace analysis consists in splitting the L-dimensional space which contains those vectors into the *signal subspace* and the *noise subspace*. The signal subspace characterizes the n sinusoids; its dimension is $p = 2n < L$. On the opposite, the noise subspace only contains noise; its dimension is $L - p$. In practice, L must be much larger than p to enhance the robustness of the algorithm.

The most interesting property of these methods for our purposes is that the estimation / subtraction of the sinusoids is not even required: the noise part can be directly obtained by projecting the subband signal onto the noise subspace. More precisely, an orthonormal basis $\boldsymbol{U}_k^S(t)$ spanning the signal subspace in the kth subband in the time window $[t-L+1, t]$ can be computed via a singular or eigenvalue decomposition of a data or a covariance matrix, or via subspace tracking methods [9]. Then, a noise vector $\boldsymbol{e}_k(t) = [e_k(t-L+1), e_k(t-L+2), \ldots, e_k(t)]^T$ can be obtained by applying the noise subspace projector $\boldsymbol{I}_L - \boldsymbol{U}_k^S(t)\,\boldsymbol{U}_k^S(t)^T$ to the subband vector $\boldsymbol{x}_k(t) = [x_k(t-L+1), x_k(t-L+2), \ldots, x_k(t)]^T$:

$$\boldsymbol{e}_k(t) = \boldsymbol{x}_k(t) - \boldsymbol{U}_k^S(t)\,\boldsymbol{U}_k^S(t)^T\boldsymbol{x}_k(t). \quad (3)$$

The noise part of the whole subband signal can be computed by an overlap-add method:

1. the analysis window $[t-L+1, t]$ is recursively time-shifted (in practice, we chose an overlap of $3L/4$),
2. the signal subspace basis $\boldsymbol{U}_k^S(t)$ is tracked by means of the *Sequential iteration EVD algorithm* [9],
3. the vector $\boldsymbol{e}_k(t)$ is computed according to Eq. (3),
4. finally, the consecutive noise vectors are multiplied by a Hanning window and added to the noise subband signal.

The overall computational cost of the subspace projection process for each analysis block is that of step 2, which is the most computationally demanding. As shown in [9], its complexity is $O(Ln(n+\log(L)))$. Note that the *Sequential iteration EVD algorthm* can be replaced by a subspace tracker of lower complexity (*e.g.* that presented in [11]).

4. SIMULATION RESULTS

The proposed tempo estimation system was tested using a data base of 54 excerpts taken from commercial recordings. These musical pieces were chosen in order to cover the following characteristics: various tempi, wide range of instruments, male/female vocals, with/without percussions. They were also selected to represent a variety of musical genres: classical music (23% of the data base); rock, modern or pop music (33%); traditional songs (12%); latin/cuban music (12%) and jazz (20%). From each of the selected recordings, an excerpt of 10 seconds long having a constant tempo was extracted and converted to a monophonic signal sampled at 16 kHz. In addition, the beat in each excerpt was meticulously manually annotated by three skilled musicians whose tempi estimations did not differ by more than 1%.

Due to the somewhat ambiguous beat definition, people tend to tap at different metrical levels. For instance, if a given track has a tempo of α BPM, some people might say that the tempo is 2α BPM, or viceversa. For evaluation purposes, the tempo estimation (T_e) provided by the algorithm is labeled as correct if it disagrees less than 5% from the manually annotated tempo (T), i.e., $0.95\,\alpha\,T < T_e < 1.05\,\alpha\,T$ with $\alpha \in \{\frac{1}{2}, 1, 2\}$. Table 1 summarizes the

filter bank	onset detect.	periodicity estim.	noise sub. prj.	correct estim.
6 (IIR)	da/dt	comb filter	NO	76 %[3]
8 (IIR)	–	YIN	NO	74 %[4]
12 (FIR)	$d \ln a/dt$	SACF	NO	87 %
8 (IIR)	”	spect. sum	NO	87 %
8 (IIR)	”	spect. prod	NO	89 %
12 (FIR)	”	SACF	YES	96 %

Table 1: *Tempo estimation performances.*

results obtained using our implementation of Scheirer [1], Paulus [4], three other methods presented in [12] and finally our new method. The spectral sum $S(e^{j\omega_n})$ and the spectral product $\Gamma(e^{j\omega_n})$ of a given signal $y(t)$ are given by Eq. (4) and Eq. (5) respectively, where $Y(e^{j\omega_n})$ stands for the DFT of $y(t)$.

$$S(e^{j\omega_n}) = \sum_{l=1}^{N} |Y(e^{jl\omega_n})|^2 \quad \text{for} \quad \omega_n < \frac{\pi}{N} \tag{4}$$

$$\Gamma(e^{j\omega_n}) = \prod_{l=1}^{N} |Y(e^{jl\omega_n})|^2 \quad \text{for} \quad \omega_n < \frac{\pi}{N} \tag{5}$$

Although the database used for testing our method has a rather limited size, simulation results indicate that it outperforms the previously mentioned algorithms. It must be pointed out that all simulations were computed using the same set of parameters. The minimal size of the analysis window (W) to achieve a success score of 96% is 6 seconds. Lower window size values reduce the system's performance for rhythmically complex signals, even though for strong beat pieces it makes no difference. The algorithm failed in 2 out of 54 excerpts, one of them is a classical music excerpt and the tempo found differed by 5.8% from the estimated one. In the second case, transients were properly detected, but the estimated tempo was erroneous because of the temporal irregularity of the onsets. During the harmonic/noise decomposition stage, several values for the number n of extracted sinusoids per subband were carefully tested and n was finally set to 5. If too many sinusoidal components are extracted, the analyzed signal is whitened and it is no longer possible to detect the attacks. If too few sinusoids are used, signal attacks are not enhanced, making onset detection more difficult. For the harmonic/noise analysis, the observation window (L) was set to 300 samples, which approximately corresponds to 19ms.

5. CONCLUSIONS

In this paper we have presented a new tempo estimation algorithm based on the harmonic/noise decomposition using noise subspace projections. Beat detection is carried out in the noise component of the signal via subband decomposition and bandwise periodicity detection. The performance of the proposed system was tested using 54 musical excerpts from several musical genres. The rate of success for tempo estimation was 96%. Future work will include an evaluation of this method using a larger database, a window size which adapts to the rhythmic complexity, and an adaptive estimation of the number of sinusoids. A real-time implementation is being considered.

[3] Our implementation of Scheirer's method.

[4] Our implementation of Paulus' method.

6. REFERENCES

[1] Scheirer, E. D., ”Tempo and Beat Analysis of Acoustic Music Signals”, *JASA*, Vol. 103: 588–601, Jan. 1998.

[2] Laroche, J., ”Estimating Tempo, Swing, and Beat Locations in Audio Recordings”, WASPAA'01, New York, USA, Oct. 2001.

[3] Goto, M., Muraoka, Y., ”Real-time Rhythm Tracking for Drumless Audio Signals”, In *Proc. of the IJCAI'97*, 1997.

[4] Paulus, J., Klapuri, A.,”Measuring The Similarity of Rhythmic Patterns”, ISMIR'02, *3rd Int. Conf. on Music Information Retrieval*, Paris, France, Oct. 2002.

[5] Seppänen, J. ”Tatum Grid Analysis of Musical Signals”, *WASPAA'01*, New York, USA, Oct. 2001.

[6] Klapuri, A., ”Sound Onset Detection by Applying Psychoacoustic Knowledge”. *Proc. ICASSP'99*, Arizona, USA, Mar. 1999.

[7] Cheveigné, A. Kawahara, H., ”YIN, a Fundamental Frequency Estimator for Speech and Music”, *JASA*, Vol. 111: No. 4, Apr. 2002.

[8] Meddis, R., O'Mard, L., ”A Unitary Model of Pitch Perception”, *JASA*, Vol. 102, Sept. 1997.

[9] Badeau, R., Boyer, R., David, B., ”EDS parametric modeling and tracking of audio signals”, *Proc. DAFx-02*, Hamburg, Germany, Sept. 2002.

[10] Keiler, F., Marchand, S., ”Survey on extraction of sinusoids in stationary sounds”, *Proc. DAFx-02*, Hamburg, Germany, Sept. 2002.

[11] Badeau, R., Richard, G., David, B., Abed-Meraim, K., ”Approximated power iterations for fast subspace tracking”, *Proc. ICASSP'03*, Apr. 2003.

[12] Alonso, M., David, B., Richard, G., ”A Study of Tempo Tracking Algorithms from Polyphonic Music Signals”. *4th COST 276 Workshop*, France, Mar 2003.