Accurate tempo estimation based on harmonic+noise decomposition

Miguel Alonso, Gael Richard, Bertrand David

Accurate tempo estimation based on harmonic+noise decomposition

Miguel Alonso, Gaël Richard and Bertrand David
GET–Télécom Paris
46 rue Barrault, 75634 Paris cedex 13, France
{miguel.alonso, gael.richard, bertrand.david}@enst.fr

Abstract

In this paper we present an innovative tempo estimation system that processes acoustic audio signals and does not use any high level musical knowledge. Our proposal relies on a harmonic plus noise decomposition of the audio signal by means of a subspace analysis method. Then, a technique to measure the degree of musical accentuation as a function of time is developed and separately applied to the harmonic and noise parts of the input signal. This is followed by a periodicity estimation block that calculates the salience of musical accents for a large number of potential periods. Next, a multi-path dynamic programming searches among all the potential periodicities for the most consistent prospects through time and finally the most energetic candidate is selected as tempo. Our proposal is validated using a manually annotated test-base containing 961 music signals from various musical genres. In addition, the performance of the algorithm under different configurations is compared. The robustness of the algorithm when processing signals of degraded quality is also measured.

I. Introduction

The continuously growing size of digital audio information increases the difficulty of its access and management, thus hampering its practical usefulness. As a consequence, the need for content-based audio data parsing, indexing and retrieval techniques to make the digital information more readily available to the user is becoming critical. It is then not surprising to observe that automatic music analysis is an increasingly active research area. One of the subjects that has attracted much attention in this field concerns the extraction of rhythmic information from music. In fact, along with harmony and melody, rhythm is an intrinsic part of the music. It is difficult to provide a rigorous universal definition, but for our needs we can quote Parncutt [1], “a musical rhythm is an acoustic sequence evoking a sensation of pulse” which refers to all possible rhythmic levels, *i.e.*, pulse rates, evoked in the mind of a listener (see figure 1). Of particular importance is the *beat*, also called *tactus* or *foot-tapping* rate, which can be interpreted as a comfortable middle point in the metrical hierarchy closely related to the human’s natural movement [2]. The concept of *phenomenal accent* has a great relevance in this context, Lerdahl and Jackendoff [3] define it as: ”the moments of musical stress in the raw signal (who) serve as cues from which the listener attempts to extrapolate a regular pattern”. In practice, we consider as phenomenal accents all the discrete events in the audio stream where there is a marked change in any of the perceived psychoacoustical properties of sound, *i.e.*, loudness, timbre and pitch.

Metrical analysis is receiving a strong interest from the community because it plays an important role in many applications: automatic rhythmic alignment of multiple instruments, channels or musical pieces; cut and paste operations in audio editing [4]; automatic musical accompaniment [5], beat driven special effects [6], [7] music transcription [8] or automatic genre classification [9].

A number of studies on metrical analysis were devoted to symbolic input usually in MIDI or other score format [10], [11]. However, since the vast majority of musical signals are available in raw or compressed audio format, a large number of recent work focus on methods that directly process the time waveform of the audio signal. As pointed out by Klapuri [8], there are three basic problems that need to be addressed in a successful metrical analysis system. First, the degree of musical stress as a function of time has to be measured. Next, the periods and phases of the underlying metrical pulses have to be estimated. Finally, the system has to choose the pulse level which corresponds to the tactus or some other specifically designated metrical level.

A large variety of approaches have already been investigated. *Histogram models* are based on the computation of the Inter-Onset Intervals (IOI) histograms from which the beat period is estimated.

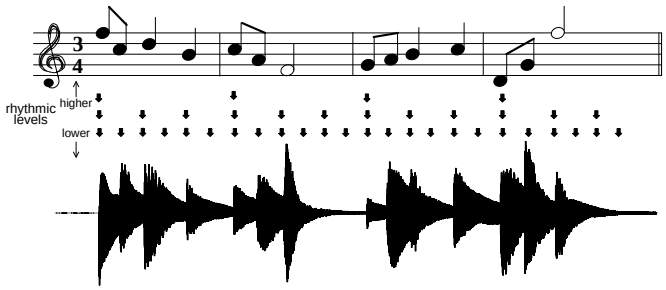

Fig. 1. Example showing how the rhythmic structure of music can be decomposed in rhythmic levels formed by equidistant pulses. There is a *double* relationship between the lowest rhythmic level and the next higher rhythmic level, on the contrary there is a *triple* relationship between the highest rhythmic level and the next lower level.

The IOI are obtained by detecting the precise location of onsets or phenomenal accents and the detectors often operate on subband signals (see for example [12], [13], [14] or [15]). The so-called *Detection Function model* does not aim at precisely extracting onset positions, but rather at obtaining a smooth profile, usually known as the "detection function", who indicates the possibility of finding an onset as a function of time. This profile is usually built from the time waveform envelope [16]. Periodicity analysis can be carried out by a bank of oscillators based on comb-filters [17], [8] or by other periodicity detectors [18], [19]. *Probabilistic models* suppose that onsets are random and exploit Bayesian approaches such as particle filtering to find beat locations [20], [21]. *Correlative* approaches have also been proposed, see [22] for a method that compares the detection function with a pulse-train signal and [23] for an autocorrelation based algorithm.

The goal of the present work is to describe a method which performs metrical analysis of acoustic music recordings at one pulsation level: the tactus. The proposed model is an extension of a previous system that was ranked first in the tempo contest of the "2^{nd} Annual Music Information Retrieval Evaluation eXchange" (MIREX) [24]. Our model includes several innovative aspects including:

- the use of a signal/noise subspaces decomposition,
- the independent processing of its deterministic (sum of sinusoids) and noise components for estimating phenomenal accents and their respective periodicity,
- the development of an efficient audio onset detector,
- the exploitation of a multi-path dynamic programming approach to highlight consistent estimates of the tactus and which allows the estimation of multiple concurrent tempi.

The paper is organized as follows. Section II describes the different elements of our algorithm, then section III presents the experimental results and compares the proposed model with two reference methods. Finally, section IV summarizes the achievements of our system and discusses possible directions for future improvements.

II. Description of the algorithm

The architecture of our tempo estimation system is provided in Figure 2. First, the audio signal is split in P subbands signals which are further decomposed into deterministic (sum of sinusoids) and noise components. From these signals, detection functions which measure in a continuous manner the degree of musical accentuation as a function of time are extracted and their periodicity is then estimated by means of several different algorithms. Next, a multi-path dynamic programming algorithm permits to robustly track through time several pulse periods from which the most persistent is chosen as the tactus. The different building blocks of our system are detailed below. Note that, throughout the rest of the paper, it is assumed that the tempo of the audio signal is stable over the duration of the observation window. In addition, we suppose that the tactus lies between 50 and 240 beats per minute (BPM).

A. Harmonic+Noise decomposition based on subspace analysis

In this part we describe a subspace analysis technique (sometimes referred to as high resolution methods) which models a signal as a sum of sinusoidal components and noise.

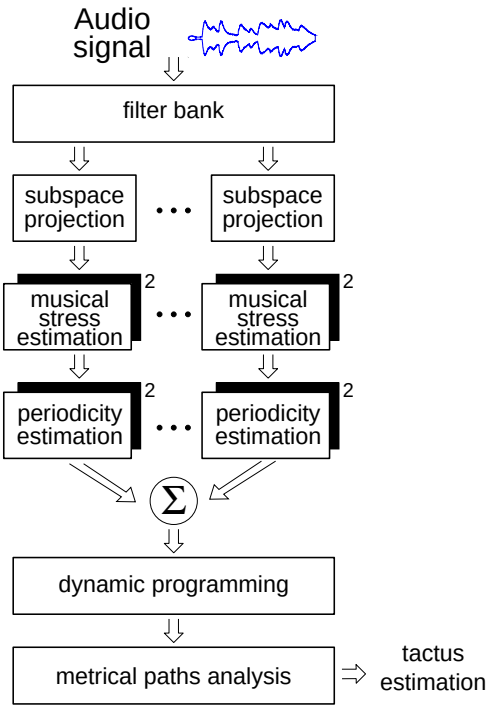

Fig. 2. Overview of the tempo estimation system.

Our main motivation to decompose the music signal is the idea of emphasizing phenomenal accents by separating them from the surrounding disturbing events, we explain this idea using an example. When processing a piano signal (percussive or plucked string sounds in general) the sinusoidal components hamper the detection of the non-stationary mechanical noise of the attack, in this case the sound of the hammer hitting the cords. Conversely, when processing a violin signal (bowed strings or wind instrument sounds in general) the non-stationary friction noise of the bow rubbing the cords hampers the detection of the sinusoidal components.

The decomposition procedure used in the present work refers to the first two blocks of the scheme presented in Figure 2 and is founded on the research carried out by Badeau *et al.* [25], [26]. Related work using such methods in the context of metrical analysis for music signals has been previously proposed in [19]. Let $x(n),\ n \in \mathbb{Z}$, be the real analyzed signal, modeled as the sum:

$$x(n) = s(n) + w(n), \tag{1}$$

where

$$s(n) = \sum_{i=1}^{2M} \alpha_i z_i^n \tag{2}$$

is referred to as the deterministic part of x. The $\alpha_i \neq 0$ are the complex amplitudes bearing magnitude and phase information and the z_i are the complex poles $z_i = e^{d_i + j2\pi f_i}$ where $f_i \in [-\frac{1}{2}, \frac{1}{2}[$ are the frequencies with $f_i \neq f_k$ for all $i \neq k$ and $d_i \in \mathbb{R}$ are the damping factors. It can be noticed that since s is a real sequence, z_i's and α_i's can be grouped in M pairs of conjugate values. Subspace analysis techniques rely on the following property of the L-dimensional data vector $\mathbf{s}(n) = [s(n-L+1), \ldots, s(n)]^T$ (with usually $2M \ll L$): it belongs to the $2M$-dimensional subspace spanned by the basis $\{\mathbf{v}(z_k)\}_{k=0,\ldots,2M-1}$, where $\mathbf{v}(z) = [1\ z\ \ldots\ z^{L-1}]^T$ is the Vandermonde vector associated with a non-zero complex number z. This subspace is the so-called *signal subspace*. As a consequence, $\mathbf{v}(z_k) \perp \text{span}(\mathbf{W}_\perp)$ where $\mathbf{W}$ denotes a $L \times 2M$ matrix spanning the signal subspace and $\mathbf{W}_\perp$ an $N \times (N - 2M)$ matrix spanning its orthogonal complement, referred to as the noise subspace. The harmonic+noise decomposition is performed by projecting the signal x respectively on the signal subspace and the noise subspace.

Initialization: $\mathbf{U}_S = \begin{bmatrix} \mathbf{I}_{2M} \\ \mathbf{0}_{(N-2M)\times 2M} \end{bmatrix}$

For each time step n iterate:
1- $\mathbf{A}(n) = \mathbf{H}(n)\mathbf{U}_S(n-1)$ fast matrix product
2- $\mathbf{A}(n) = \mathbf{U}_S(n)\mathbf{R}(n)$ skinny QR factorization

TABLE I
SEQUENTIAL ITERATION EVD ALGORITHM.

Let the symmetric $L \times L$ real Hankel matrix $\mathbf{H}_s$ be the data matrix:

$$\mathbf{H}_s = \begin{bmatrix} s(0) & s(1) & \cdots & s(L-1) \\ s(1) & s(2) & \cdots & s(L) \\ \vdots & \vdots & \ddots & \vdots \\ s(L-1) & s(L) & \cdots & s(N-1) \end{bmatrix}, \tag{3}$$

where $N = 2L-1$, with $2M \leq L$. Since each column of $\mathbf{H}_s$ belongs to the same $2M$-dimensional subspace, the matrix is of rank $2M$ and thus is rank deficient. Its eigenvalue decomposition (EVD) yields

$$\mathbf{H}_s = \mathbf{U}\mathbf{\Lambda}_s\mathbf{U}^H \tag{4}$$

where $\mathbf{U}$ is an orthonormal matrix, $\mathbf{\Lambda}_s$ is the $L \times L$ diagonal matrix of the eigenvalues, $L-2M$ of which are zeros. $\mathbf{U}^H$ denotes the Hermitian transpose of $\mathbf{U}$. The $2M$-dimensional space spanned by the columns of $\mathbf{U}$ corresponding to the non-zero entries of $\mathbf{\Lambda}_s$ is the signal subspace.

Because of the surrounding additive white noise $\mathbf{H}_x$ is full rank and the signal subspace, $\mathbf{U}_S$, is formed by the $2M$-dominant eigenvectors of $\mathbf{H}_x$, *i.e.*, the column of $\mathbf{U}$ associated to the $2M$ eigenvalues having the highest magnitudes.

In practice, we observe the noisy sequence $x(n)$ and its harmonic part can be obtained by projecting $x(n)$ onto its signal subspace as follows:

$$\mathbf{s} = \mathbf{U}_S\mathbf{U}_S^H\mathbf{x} \tag{5}$$

A remarkable property of this method is that for calculating the noise part of the signal, the estimation and subtraction of the sinusoids is not required explicitly. The noise is obtained by projecting $x(n)$ onto the noise subspace:

$$\mathbf{w} = \mathbf{x} - \mathbf{s} = (\mathbf{I} - \mathbf{U}_S\mathbf{U}_S^H)\mathbf{x}. \tag{6}$$

Subspace tracking. Since the harmonic plus noise decomposition of $x(n)$ involves the calculation of one EVD of the data matrix $\mathbf{H}_x$ at every time step, decomposing the whole signal would require a highly demanding computational burden. In order to reduce this cost, there exist adaptive methods that avoid the computation of the EVD [27], a survey of such methods can be found in [26]. For the present work, we use an iterative algorithm called *sequential iteration* [25], shown in Table I. Assuming that it converges faster than the variations of the signal subspace, the algorithm operation involves two auxiliary matrices at every time step $\mathbf{A}(n)$ and $\mathbf{R}(n)$, in addition of a skinny QR factorization. The harmonic and noise parts of the whole signal $x(n)$ can be computed by means of an overlap-add method:

1) the analysis window is recursively time-shifted. In practice, we choose an overlap of $3L/4$,
2) the signal subspace $\mathbf{U}_S$ is tracked by means of the previously mentioned sequential iteration algorithm presented in Table I.
3) the harmonic, $\mathbf{s}$, and noise, $\mathbf{w}$, vectors are computed according to Eq. (5) and Eq. (6),
4) finally, consecutive harmonic and noise vectors are multiplied by a Hann window and respectively added to the harmonic and noise parts of the signal.

The overall computational complexity of the harmonic plus noise decomposition for each analysis block is that of step 2, which is the most computationally demanding task of the whole metrical analysis system. Its complexity is $O(L\,n(n+\log(L))$.

Subspace analysis methods rely on two principles. From one part they assume that the noise is white and secondly, that the order of the model (number of sinusoids) is known in advance. Both of these premises are not usually satisfied in most applications.

A practical remedy to overcome the colored noise problem consists in using a pre-accentuation filter[1] and in separating the signal in frequency bands, which has the effect of leading to a (locally) whiter noise in each channel. The input signal $x(n)$ is decomposed into $P = 8$ uniform subband signals $x_p(n)$, where $p = 0, \ldots, P-1$. Subband decomposition is carried out using a maximally decimated cosine modulated filter bank [28], where the prototype filter is implemented as a 150^{th} order FIR filter with 80 dB of rejection in the stop band. Using such a highly selective filter is relevant because subspace projection techniques are very sensitive to spurious sinusoids.

Estimating the exact number of sinusoids present in a given signal is a considerably difficult task and a large effort has been devoted to this problem, for instance [29] [30]. For our application we decided to slightly overestimate the model order since according to Badeau [26, page 54] it has a small impact in the algorithm performance compared to an underestimation. Another important advantage of the bandwise processing approach is that there are less sinusoids per subband (compared to the full band signal) which allows at the same time to reduce the overall computational complexity, *i.e.*, we deal with more matrices but P-times smaller in size.

In this way, further processing in the subbands is the same for all frequency channels. The output of the decomposition stage consists in two signals: $s_p(n)$ carrying the harmonic and $w_p(n)$ the noise part of $x_p(n)$.

B. Calculation of a musical stress profile

The harmonic+noise decomposition previously described can be seen as a front-end that performs "signal conditioning", in this case it consists in decomposing the input signal in several harmonic and noise components prior to rhythmic processing.

In the metrical analysis community there exists an implicit consensus about decomposing the music signal in subbands prior to conducting rhythm analysis. According to experiments carried out by Scheirer [17], there exists no optimal decomposition since many subband layouts lead to comparable satisfactory results. In addition, he argues that a "psychoacoustic simplification" consisting in a simple envelope extraction in a number of subbands is sufficient to extract pulse and meter information from music signals. The tempo estimation system herein proposed is built upon this principle.

The concept of phenomenal accent as a discrete sound event plays a fundamental role in metrical analysis. Humans hear them in a hierarchical structure, that is, a phenomenal accent is related to a motif, several motifs are clustered into a pattern and a musical piece is formed of several patterns that may be different or not. In the present work, we attempt to be acute (in a computational sense) to the physical events in an audio signal related to the moments of musical stress, such as magnitude changes, harmonic changes and pitch leaps. That is, acoustic effects that can be heard and are musically relevant for the listener. The attribute of being sensitive to these events does not necessarily implies the need of a specific algorithm for detecting harmonic or pitch changes, but solely a method which reacts to variations in these characteristics.

In practice, calculating a profile of the musical stress present in a music signal as a function of time is intimately related to the task of detecting onsets. Robust onset detection for a wide range of music signals has proven to be a difficult task. In [31] Bello provides a survey of the most commonly used methods. While we propose an approach that exploits previous research [16], [22] as a starting point, it significantly improves the calculation of the Spectral Energy Flux (SEF) or spectral difference [32]. See Figure 3 for an overview of the proposed method. As in the previous section, the algorithm will be presented for a single subband case and only for the harmonic component $s_p(n)$, since the same procedure is followed for the noise part $w_p(n)$ and the rest of the subbands.

Spectral energy flux. The method that we present resides on the general assumption that the appearance of an onset in an audio stream leads to a variation in the signal's frequency content. For example, in the case of a violin producing pitched notes, the resulting signal will have a strong fundamental frequency that leaps in time as well as the related harmonic components at integer

[1]Since the power spectral density of audio signals is a decreasing function of frequency, the use of a pre-accentuation filter that tends to flatten this global trend is necessary. In our implementation we use the same filter as in [26], that is: $G(z) = 1 - 0.98z^{-1}$

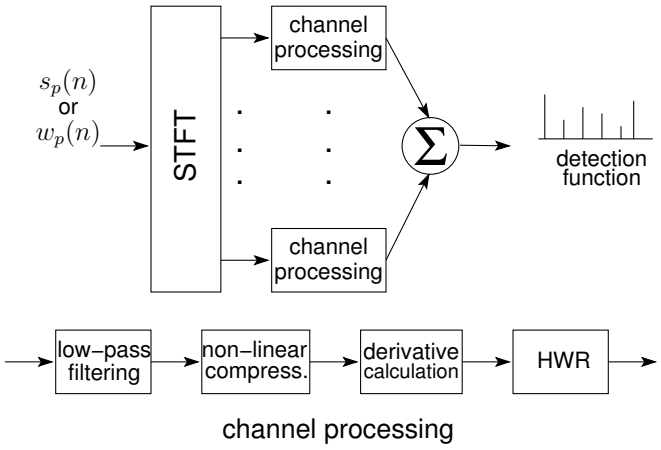

Fig. 3. Overview of the system to estimate musical stress.

multiples of the fundamental attenuating as frequency increases. In the case of a percussive instrument, the resulting signal will tend to have sharp energy boosts. The harmonic component $s_p(n)$ is analyzed using the STFT, leading to

$$\tilde{S}_p(m,k) = \sum_{n=-\infty}^{\infty} \mathrm{w}(Mm-n)s_p(n)e^{-j\frac{2\pi}{N}kn} \tag{7}$$

where $\mathrm{w}(n)$ is a finite-length sliding window, M the hop size, m the time (frame) index and $k = 0,\ldots,N-1$ the frequency channel (bin) index. To detect the above mentioned variations in the frequency content of the audio signal, previous methods have proposed the calculation of the derivative of $\tilde{S}_p(m,k)$ with respect to time

$$E_p(l,k) = \sum_m h(l-m)\, G_p(m,k) \tag{8}$$

and where $E_p(l,k)$ is known as the Spectral Energy Flux (SEF), $h(m)$ is an approximation to an ideal differentiator

$$H(e^{j2\pi f}) \simeq j2\pi f \tag{9}$$

and

$$G_p(m,k) = \mathcal{F}\{|\tilde{S}_p(m,k)|\} \tag{10}$$

is a transformation that accentuates some of the psychoacoustically relevant properties of $\tilde{S}_p(m,k)$.

In solving many physical problems by means of numerical methods, it is a challenge to seek derivatives of functions given in discrete points. For example, in [16], [22] authors propose a first order difference with $h = [1,\, -1]$, which is a rough approximation to an ideal differentiator. In this paper, we use a differentiator filter $h(m)$ of order $2L$ based on the formulæ for central differentiation developed by Dvornikov in [33] which provides a much closer approximation to (9). Other efficient differentiator filters can be used providing comparable results, for instance, FIR filters obtained by the Remez method [34]. The underlying principle of the proposed digital differentiator is the calculation of an interpolating polynomial of order $2L$ passing through $2L+1$ discrete points which is used to find the derivative approximation. A comprehensive description of the method and its accuracy to approximate Equation (9) can be found in [33]. The analytical expression to compute the first L coefficients of an antisymmetric FIR differentiator is given by $g(i) = \frac{1}{i\alpha(i)}$ with

$$\alpha(i) = \prod_{\substack{j=1 \\ j\neq i}}^{L} \left(1 - \frac{i^2}{j^2}\right) \tag{11}$$

and $i = 1,\ldots,L$. The coefficients of $h(m)$ are given by

$$h = [-g(L),\, \ldots,\, 0,\, \ldots,\, g(L)]. \tag{12}$$

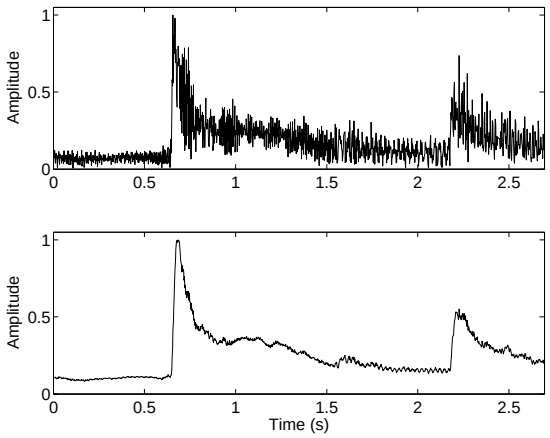

Fig. 4. The smoothing effect of the energy integration function emphasizes signal attacks but masks rapid modulations. The image shows a *pitched* frequency channel corresponding to piano signal (upper part) before smoothing and (lower part) after smoothing.

In our proposal, the transformation $G(m,k)$ calculates a perceptually plausible power envelope for frequency channel k and is formed of two steps. First, psychoacoustic research on computational models of mechanical to neural transduction [35] shows that the auditory nerve adaptation response following a sudden stimulus change can be characterized as the sum of two exponential decay functions:

$$\phi(m) = \alpha e^{-m/T_1} + \beta e^{-m/T_2} \quad \text{for} \quad m \geq 0 \tag{13}$$

formed by a rapid decline component with time constant (T_1) in the order of 10 ms and a slower short-term decline with a time constant (T_2) in the region of 70 ms. This adaptation function performs energy integration, emphasizing the most recent stimulus but masking rapid modulations. From a signal processing standpoint, this can be viewed as two smoothing low-pass filters whose impulse response has a discontinuity that preserves edge sharpness and avoids dulling signal attacks. In practice, the smoothing window is implemented as a 2nd-order IIR filter with z-transform

$$\Phi(z) = \frac{\alpha + \beta - (\alpha z_2 + \beta z_1\, z^{-1})}{1 - (z_1 + z_2)z^{-1} + z_1 z_2\, z^{-2}}. \tag{14}$$

where $T_1 = 15$ ms, $T_2 = 75$ ms, $\alpha = 1$, $\beta = 5$, $z_1 = e^{-1/T_1}$ and $z_2 = e^{-1/T_2}$. Figure 4 shows the role of the energy integration function after convolving it with a pitched channel of a signal's spectrogram representation.

The second part of the envelope extraction consists in a logarithmic compression. This operation has also a perceptual relevance since the logarithmic difference function gives the amount of change in a signal's intensity in relation to its level, that is

$$\frac{d}{dt}\log I(t) = \frac{\Delta I(t)}{I(t)}. \tag{15}$$

This means that the same amount of increase is more prominent in a quiet signal [16], [36].

In practice, the algorithm implementation is straightforward, and is carried out as presented in Figure 3. The STFT in Equation (7) is computed using an N point fast Fourier transform (FFT). The absolute value of every frequency channel, $|\tilde{S}(m,k)|$ is convolved with $\phi(m)$. The smoothing operation is followed by a logarithmic compression. The resulting $G(m,k)$ is given by

$$G(m,k) = \log_{10}\left(\sum_i |\tilde{S}(i,k)|\phi(m-i)\right). \tag{16}$$

At those time instants where the frequency content of $s_p(n)$ changes and new frequency components appear, $E(l,k)$ exhibits positive peaks whose amplitude is proportional to the energy and rate of change of the new components. In a similar way, when frequency components disappear from $s_p(n)$, the SEF exhibits negative peaks, marking the *offset* of a musical event. Since we are only interested in onsets, we apply a half-wave rectification (HWR) to $E(l,k)$, *i.e.*, only positive values are taken into account.

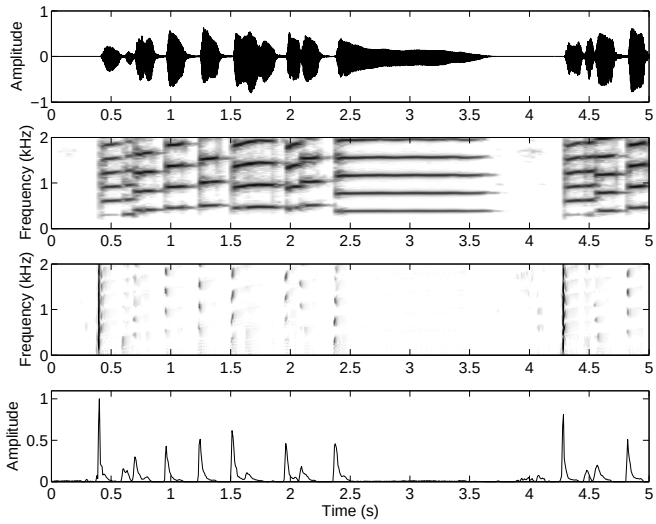

Fig. 5. Trumpet signal example, see text for a description. From top to bottom: harmonic part waveform, spectrogram representation, the corresponding spectral flux $E(l,k)$ and the detection function $v(l)$.

To find a global stationarity profile $v(l)$, better known as the *detection function*, contributions from all channels are integrated across frequency

$$v(l) = \sum_{\substack{k \\ E(l,k)>0}} E(l,k). \tag{17}$$

$v(l)$ displays sharp peaks at transients and note onsets, those instants where the positive energy flux is large. Figure 5 shows an example for a trumpet signal. From top to bottom: waveform of the harmonic part for the subband $s_0(n)$; the respective STFT modulus, highlighting the signal's harmonic structure; SEF $E(l,k)$, dotted vertical edges indicate the regions where the SEF is large; the bottom part presents the detection function $v(l)$, onset instants and intensity are indicated by peaks location and height respectively.

The output of the phenomenal accent detection stage is formed of two signals per subband: the harmonic part detection function $v_p^s(l)$, and the noise part detection function $v_p^w(l)$.

C. Periodicity estimation

The basic constituents of the comb-like detection functions $v_p^s(l)$ and $v_p^w(l)$ are pulsations representing the underlying metrical levels. The next step consists in estimating the periodicities embedded in those pulsations. This analysis takes place at a subband level for both harmonic and noise parts. As briefly mentioned in Section I, many periodicity estimation algorithms have been proposed to accomplish this task. In the present work, we test three different methods widely used in pitch determination techniques: the spectral sum, the spectral product and the autocorrelation function. The procedure described below is repeated $2p$ times to account for the harmonic and noise parts in all subbands. In this stage, no decisions about the pulse frequencies present in $v_p(l)$ are taken, but only a measure of the degree of periodicity present in the signal is calculated. First, $v_p(l)$ is decomposed into contiguous frames g_n with $n = 0, \ldots, N-1$ of length ℓ and an overlapping of ρ samples, as shown in Figure 6. Then, a periodicity analysis of every frame is carried out producing a signal r_n of length K samples generated by any of the three methods explained below:

1) Spectral sum: The spectral sum (SS) method relies on the assumption that the spectrum of the analyzed signal is formed of strong harmonics located at integer multiples of its fundamental frequency. To find periodicities, the power spectrum of g_n, *i.e.*, $|G_n(e^{j2\pi f})|$, is compressed by a factor λ, then the obtained spectra are added, leading to a reinforced fundamental. For normalized frequency, this is given by

$$r_n = \sum_{\lambda=1}^{\Lambda} |G_n(e^{j2\pi\lambda f})|^2 \quad \text{for} \quad f < \frac{1}{2\Lambda} \tag{18}$$

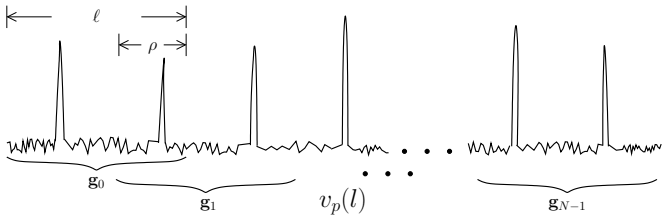

Fig. 6. Decomposition of $v_p(l)$ into contiguous overlapping windows g_n.

where Λ is the upper compression limit that ensures that half the sampling frequency is not exceeded. The spectral sum corresponds to the maximum likelihood solution of the underlying estimation problem.

2) Spectral product: The spectral product (SP) method is quite similar to the above mentioned SS, the only difference consists in substituting the sum by a product, that is

$$r_n = \prod_{\lambda=1}^{\Lambda} |G_n(e^{j2\pi\lambda f})|^2 \quad \text{for} \quad f < \frac{1}{2\Lambda} \tag{19}$$

3) Autocorrelation: The biased deterministic autocorrelation (AC) of g_n

$$r_n = \frac{1}{\ell} \sum_{l} g_n(l+\tau) g_n(l). \tag{20}$$

Data fusion. Once all r_n have been calculated, they are fused in a two step process. First, every r_n from the harmonic and noise parts is normalized by its largest value and weighted by a peakness coefficient[2] c_n calculated over the corresponding g_n. In this way, we penalize flat windows g_n (bearing little information) by a low weighting coefficient $c_n \approx 0$. On the opposite side, a peaky window g_n leads to $c_n \approx 1$. The second step consists in adding information from all subbands coming from both harmonic and noise parts:

$$\gamma_n = \frac{1}{2P} \sum_{p=1}^{P} c_{n,p}^{s} r_{n,p}^{s} + \frac{1}{2P} \sum_{p=1}^{P} c_{n,p}^{w} r_{n,p}^{w} \tag{21}$$

where the superscript s and w on the right side indicate the harmonic and noise part respectively. Since this frame process is repeated N times, then all the resulting γ_n are arranged as column vectors ($\boldsymbol{\gamma}_n$) to form a periodicity matrix $\mathbf{\Gamma}$ of size $K \times N$ as follows

$$\mathbf{\Gamma} = [\boldsymbol{\gamma}_0 \ \boldsymbol{\gamma}_1 \ \cdots \ \boldsymbol{\gamma}_{N-1}]. \tag{22}$$

$\mathbf{\Gamma}$ can be seen as a time–frequency representation of the pulsations present in $x(n)$, since rows exhibit the degree of periodicity at different frequencies, while columns indicate their course through time.

D. Finding and tracking the best periodicity paths

At this point of the analysis, we have a series of metrical level candidates whose salience over time is registered in the columns of $\mathbf{\Gamma}$. The next stage consists of parsing through the successive columns to find at each time instant n the best candidates and thus track their evolution. Dynamic programming (DP) is a technique that has been extensively used to solve this kind of sequential decision problems, details about its implementation can be found in [37]. In addition, it has also been proposed for metrical analysis [22], [38]. At each time frame n there exists K potential path candidates called $\Gamma_{(n,k)}$. The DP solves this combinatorial optimization problem by examining all possible combinations of the $\Gamma_{(n,k)}$ in an iterative and rational fashion. Then, a path is formed by concatenating a series ψ_n of selected candidates from each frame: the $\Gamma_{(n,\psi_n)}$. The DP procedure iteratively defines a score $\mathcal{S}_{(n,k)}$ for a path arriving at candidate $\Gamma_{(n,k)}$ and this score is a function of

[2]In the present work we use as peakness measure $c = 1 - \phi$, where $\phi = \left(\prod_{l=1}^{\ell} g(l)\right)^{1/\ell} / \left(1/\ell \sum_{l=1}^{\ell} g(l)\right)$. Since ϕ (the ratio of the geometric mean to the arithmetic mean) is a flatness measure bounded to the region $0 < \phi \leq 1$, when $c \to 1$ it means that $g(l)$ has a peaked-shape. On the contrary, if $c \to 0$, means that $g(l)$ has a flat-shape.

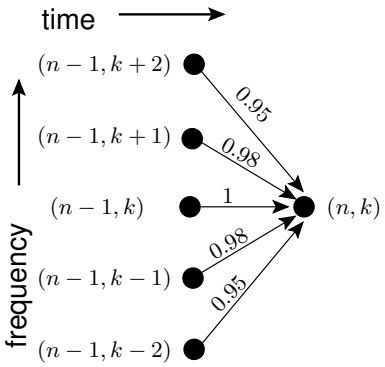

Fig. 7. Dynamic programming local constraint for path tracking.

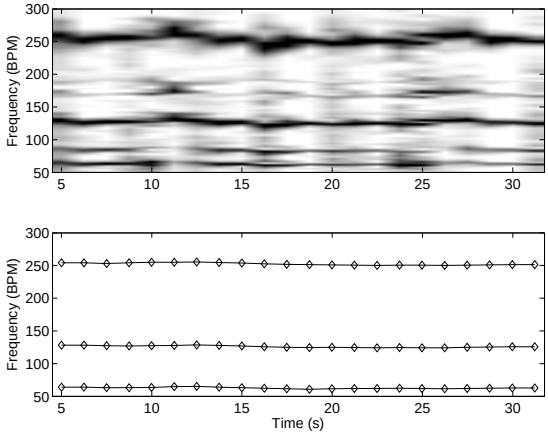

Fig. 8. Tracking of three most salient periodicity paths for Mozart's *Rondo Alla Turca*. The relationship among them is 1:2:4

three parameters: the score of the path at the previous frame $\mathcal{S}_{(n-1,\psi_{n-1})}$, where $\psi_{(n-1)}$ represents the candidate through which the path comes from time $n-1$; the periodicity salience of the candidate under analysis $\Gamma_{(n,k)}$; and a transition penalty, also called local constraint, $D_{(\psi_{n-1},k)}$ who deprecates the score of a transition from candidate ψ_{n-1} at time $n-1$ to candidate k at time n according to the rule shown in Figure 7. These three parameters are related in the following way:

$$\mathcal{S}_{(n,k)} = \mathcal{S}_{(n-1,\psi_{n-1})} D_{(\psi_{n-1},k)} + \Gamma_{(n,k)}. \tag{23}$$

The transition-penalty rule relies on the assumption that in common music, metrical levels generally vary slowly in time. In our implementation, a transition in the vertical axis of one position corresponds to about 1 BPM (the exact value depends on the method used to estimate the periodicity). Thus, the DP smoothes the metrical level paths and avoids abrupt transitions. In addition, the DP stage has been designed such that paths sharing segments or being too close ($<$ 10 BPM) to more energetic paths are pruned. Figure 8 shows an example of the DP performance, in the upper part can be seen an image of the time–frequency matrix $\mathbf{\Gamma}$ for Mozart's piece *Rondo Alla Turca* showing in black shades the salience. In the lower part are shown the three most salient paths obtained by the DP algorithm and representing metrical levels related as 1:2:4. To estimate the tactus, the path with highest energy (*i.e.*, the most persistent through time) is selected and the average of its values is computed. If a second most salient periodicity is required (for example, as demanded in the MIREX'05 "tempo extraction contest") the average of the second most energetic path obtained by the DP algorithm is provided as secondary tactus.

III. Performance analysis

In this section, we present the evaluation of the proposed system. Its performance under different situations is also addressed, along with a comparison to another reference method. Note that the

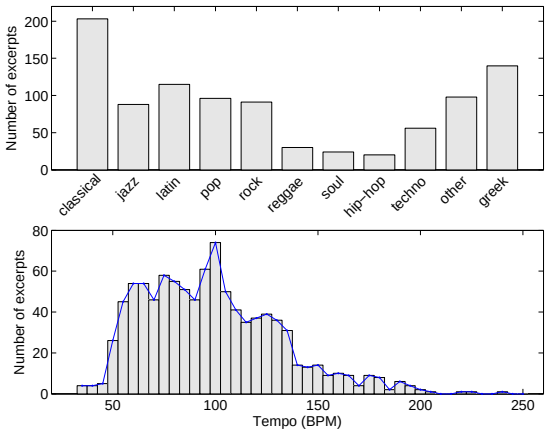

Fig. 9. Dataset information. On Top the genre distribution in the database and in the bottom ground-truth tempi distribution.

tempo estimation system includes beat-tracking capabilities, although this task is not evaluated in the present paper.

A. Test data and evaluation methodology

The proposed system was evaluated using a corpus of 961 musical excerpts taken from two different datasets. Approximately 56% of the data comes from the authors' private collection, while the rest is the *song excerpts* part of the ISMIR'04 "tempo induction contest" [39] for which data and annotations are freely available. The musical genres and tempi distribution of the database used to carry out the tests are presented in Figure 9. Genre categories were selected according to those of *Amazon.com*®. To construct both databases, musical excerpts of 20 s with a relatively constant tempo were extracted from commercial CD recordings, converted to monophonic format and downsampled at 16 kHz with 16 bits resolution. In the authors' private database, each excerpt was meticulously manually annotated by three skilled musicians who tapped along with the music while the tapping signal was being recorded. The *ground-truth* was computed in a two step process. First, the median of the inter-beat intervals was calculated. Then, concording annotations from different annotators were directly averaged, while annotations differing by an integer multiple were normalized in order to agree with the majority before being averaged. If no consensus was found the excerpt was rejected. The *song excerpts* database was annotated by a professional musician who placed beat marks on song excerpts and the ground-truth was computed as the median of the inter-beat intervals [40].

Quantitative evaluation of metrical analysis systems is an open issue. Appropriate methodologies have been proposed [41], [42], however they rely on an arduous or extremely time-consuming annotation process to obtain the ground-truth. Due to such limitations in the annotated data, the quantitative evaluation of the proposed system was confined to the task of estimating the scalar value of the tactus (in BPM) of a given excerpt, instead of an exhaustive evaluation at several metrical levels involving beat-rates and phase locations. A first step towards benchmarking metrical analysis systems has been proposed in [40]. In a similar way, during our evaluation two metrics are used:

- *Accuracy 1*: the tactus estimation must lie within a 5% precision window of the ground-truth tactus,
- *Accuracy 2*: the tactus estimation must lie within a 5% precision window of the ground-truth tactus or half, double, three times or one-third of the ground-truth tactus.

The reason for using the second metric is motivated by the fact that the ground-truth used during the evaluation does not necessarily represent the metrical level that most of human listeners would choose [40]. This is a widespread assumption found among metrical systems evaluations.

B. Experimental results

1) Effect of window length and overlap: it is interesting to know if the combination of the three periodicity algorithms that we use (SS, SP and AC) would reach a score higher than individual entries.

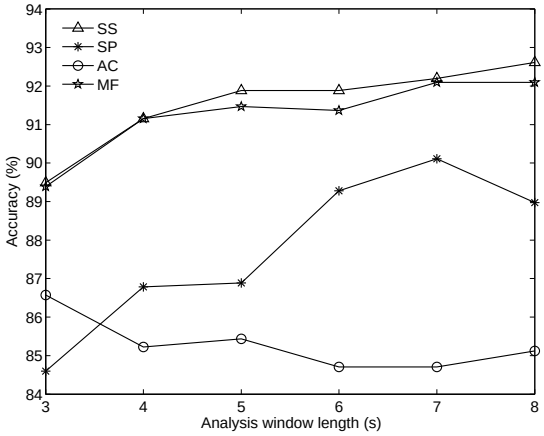

Fig. 10. On the influence of window length.

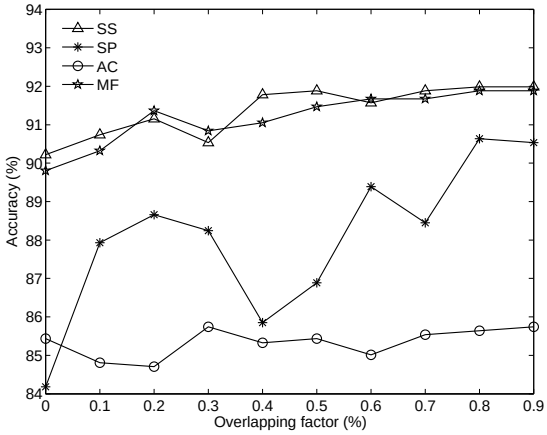

Fig. 11. On the influence of the window overlap.

For this reason we created a fourth entrant called *Method Fusion* (MF) that combines results from the three other methods using a majority rule. If there exists no agreement between methods, preference was given to the SS. To measure the impact of the window length ℓ, the overlapping was fixed to $\rho = 0.5\ell$. Then, several values of ℓ were tested as shown in Figure 10. For the spectral methods a performance gain is obtained as ℓ increases. This improvement is especially important for the approach based on the SP. In the case of the AC, increasing ℓ was counterproductive, since it slightly degraded the performance probably due to the influence of the spurious peaks in $v_p(l)$. There exists a trade off between window length and adaptability to rhythmic fluctuations. From Figure 10 it can be seen that accuracy for the SS and MF methods has practically reached its maximum when $\ell = 5$ s. We then study the overlapping ρ parameter influence on the overall performance for a fixed window length ($\ell = 5$ s). Figure 11 clearly shows that introducing this redundancy in the time–frequency matrix $\mathbf{\Gamma}$ yields a significant gain in performance for the SS, SP and MF methods, this can be explained by the fact that the DP stage has a larger data horizon and adapts better to metrical levels paths. For the AC method, varying ρ does not seem to have a significant effect in the results. As in the ℓ case, large ρ values bring a loss in adaptability. We fixed the overlapping to $\rho = 0.6\ell$, since it provides a "good" trade-off between accuracy and tracking capability. Hereafter, all results will be computed using $\ell = 5$ s and $\rho = 0.6\ell$.

2) Performance per genre: Figure 12 presents the algorithms' performance in the form of bars showing accuracy *vs.* musical genre, these results were calculated using the *Accuracy 1* criterion. Figure 13 presents the algorithms' performance but this time using the *Accuracy 2* criterion. Results are in general considered satisfactory. With the only exception of greek music, for all genres at least one of the periodicity methods obtained a score above 90%. For the reggae, soul and hip-hop genres

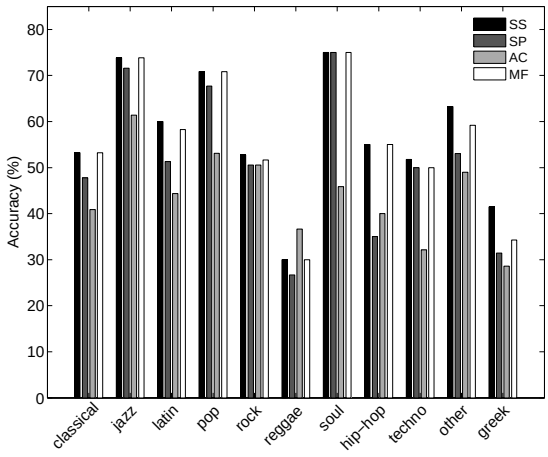

Fig. 12. Operation point (5 s, 60% overlap) performance, *Accuracy 1*.

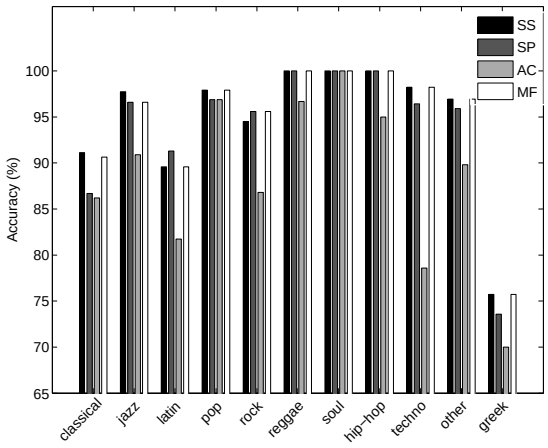

Fig. 13. Operation point (5 s, 60% overlap) performance, *Accuracy 2*.

in some cases even a success rate of 100% was obtained (under the *Accuracy 2* criterion), although such results must be taken with cautious optimism since these genres are not particularly difficult and their representation in the dataset is rather limited, as shown in Figure 9. For enhancement purposes, it is perhaps more interesting to analyze the instances where the algorithm failed. For the classical genre, the cases where the algorithms failed are mostly related to smooth onsets (usually in string passages) that are not detected. In some excerpts a wrong metrical level was chosen (for example 2/3 of the tempo). In the jazz case, most failures are related to poly-rhythmic excerpts where the tactus found by the algorithm differed from the one selected by the annotators. For the latin, pop, rock, "other" and greek genres, the large majority of the errors are found in excerpts with a strong speech foreground or having large chorus regions, both incorrectly managed by the onset detection stage. For the greek genre, poly-rhythmic excerpts with a peculiar time-signature are often the cause of a wrong detection. In techno music, some digital sound effects lead to false onsets.

3) Impact of the harmonic+noise decomposition: A natural question arises when we inquire about the influence of the harmonic plus noise decomposition in the system's performance. To answer it, the proposed method has been slightly modified and the *subspace projection* block presented in Figure 2 has been bypassed. This modified approach is based on a previous system that has been compared to other state of the art algorithms and was ranked first in the "2^{nd} Annual Music Information Retrieval Evaluation eXchange" (MIREX) in the "Audio Tempo Extraction" category. Evaluation details and results are available on-line [24], [43]. Besides, we decided to assess the contribution of the harmonic plus noise decomposition proposed in section II-A (EVD H+N) by comparing it to a more common approach based on the STFT (FFT H+N). The principle used to perform this decomposition is close to that proposed by [44]. In addition, we compared the above mentioned system variations to the well-

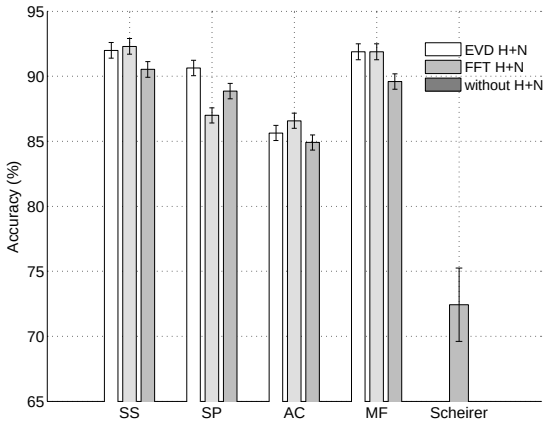

Fig. 14. Algorithm comparison to see the influence of the H+N decomposition. The error-bar indicates the 95% confidence interval calculated as $1.96\sqrt{\frac{pq}{N}}$ where p corresponds to the accuracy (in the $[0\ 1]$ range) of the algorithm under analysis, q is computed by $q = 1 - p$ and N is the total number of excerpts under analysis [45, page 47].

known classical method proposed by Scheirer[3] [17]. A small modification of Scheirer's algorithm output was carried out, since it was conceived to produce a set of beat times rather than an overall scalar estimate of the tactus.

The accuracies of the algorithms can be seen in Figure 14. While the proposed system (EVD H+N) attained a maximum score of 92.0%, it was slightly outperformed by its variation based on the STFT decomposition (FFT H+N), who obtained 92.3% of accuracy (both under the SS method). All tests showed better performance for the H+N based approaches, with the exception of the STFT decomposition (FFT H+N) when combined with the SP periodicity estimation method. The results shown in Figure 14 suggest that the statistical significance in the accuracy between carrying out a H+N decomposition or not, depends on the method used. While the SS and MF show a small but consistent improvement, the SP and AC fail to present the H+N decomposition as statistically advantageous. Nevertheless, a general trend indicating a better performance is perceived.

After taking a closer look at the improvement obtained by using the H+N decomposition, we can see that it is mainly formed of excerpts containing weak attacks such as bowed-string and wind instruments, and to a lesser extent of signals with a rather clear rhythm but with a salient speech foreground (vocals). When we examined the excerpts for which none of the algorithms succeeded, we found practically the same kind of signals: bowed-strings with large vibratos and weak attacks, orchestral pieces and signals with a strong speech foreground. In fact, the weakness of the algorithm lies in the musical stress estimation module. This can be seen as a single problem formed of two different facets:

- the incapability of detecting soft attacks mainly seen in classical pieces, while visual inspecting the set of detection functions we noticed that true attacks do not surpass the noise level;
- the presence of too many false attacks in the detection function, mainly provoked by the appearance of local frequency variations seen in vibratos and speech signals.

Both kinds of malfunctions produce an erroneous periodicity profile and consequently a wrong tempo estimation.

As can be seen from Figure 14, the majority rule combination of the three periodicity estimation methods (MF) did not obtain the best performance. Since the SS has the higher score among all methods proposed, it will be the only one considered in the next part of the analysis.

4) Robustness to signal degradation: In order to evaluate robustness to signal degradation, we used the scenario suggested by Gouyon *et al.* [40] with minor modifications: every excerpt was downsampled, GSM[4] encoded/decoded, up-sampled at 16 kHz, band-pass filtered in the 500–4000 Hz range, reverberation with a delay of one second was added and finally corrupted by white Gaussian

[3]This version of Scheirer's algorithm was ported from the Dec Alpha platform to GNU/Linux by Anssi Klapuri.

[4]Based on the digital speech codec GSM 06.10 "Regular Pulse Excitation Long-Term Predictor" (RPE-LTP) compressing at 13 kbps.

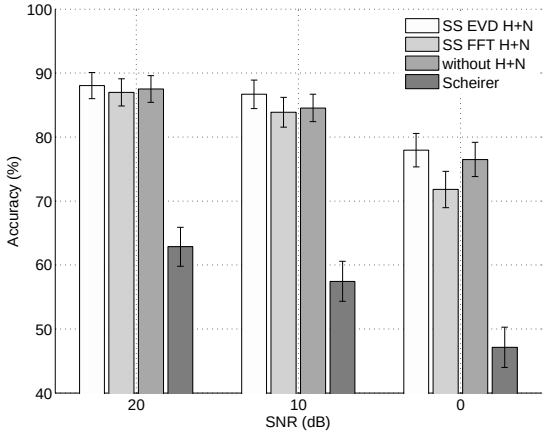

Fig. 15. Robustness to signal degradation. The EVD H+N algorithm displays the highest strength to signal distortion.

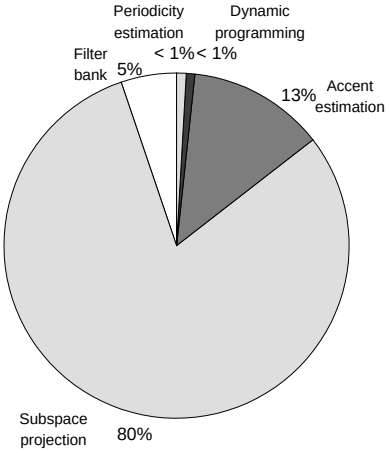

Fig. 16. Computational cost of the tempo estimation system. The total processing time required for analyzing a 20 s musical excerpt time is 23.248 s.

noise at three different SNRs. The performance of the evaluated systems are presented in Figure 15. While the EVD H+N version displays an outstanding robustness to signal distortion, its counterpart FFT H+N shows to be more sensitive, even than the non-decomposition approach. This fact becomes more evident as the SNR reduces, however the interest of the H+N approach for noise robustness is questionable in this case since the difference is not statistically significant. The EVD H+N robustness to signal degradation has been previously exploited in the literature as a denoising tool for speech signals in automotive applications [46], [47]. As long as the SNR is high enough to guarantee that the $2M$-dominant eigenvectors of $\mathbf{H}_x$ (see Section II-A) effectively correspond to the audio signal, the harmonic part ($s_p(n)$) will be noise free. If the SNR is further reduced, spurious components will be detected among the dominant eigenvectors, as a result the harmonic part will be corrupted. Figure 15 also shows Scheirer's algorithm robustness to signal distorsion.

5) Computational cost: A key attribute of any tempo estimation system is its computational complexity. Since we implemented our algorithm under Matlab© 6.5.1 (R13) and we use a number of built-in functions, a meticulous evaluation appears to be rather complicated. The approach we adopted to estimate the burden is not the most infallible, but it is the most straightforward yet providing a tangible opinion about the true complexity. We measured the time it takes to the EVD H+N algorithm to process a 20 s excerpt taken from the test-base. Figure 16 shows the percentage consumption per analysis block and the total processing time was 23.248 s. This figures were obtained using a Pentium 4 machine running at 2.4 GHz with 512 MB of memory under Debian GNU/Linux 3.1 (Sarge). The subspace projection stage is by far the most time consuming block.

IV. Conclusions

In this paper we have presented a system that successfully analyzes acoustic music recordings in order to extract tactus information. The proposed method was validated using a large dataset containing 961 instances covering several musical genres. Without requiring any high-level music information, our system shows that a good accuracy can be obtained using a common system configuration and the same parameter set. Moreover, our results indicate that decomposing the audio signal into harmonic and noise parts prior to rhythm analysis yields a small but consistent improvement in performance and proved to be robust to signal distortion. The major drawback of the system is that this accuracy increase was obtained at the expense of a high computational cost. It must be remarked that the combination of the system components (harmonic and noise) is rather crude and this may explain that only a small improvement in performance is obtained. Further work should be dedicated to the elaboration of improved fusion strategies. We have also presented a technique to estimate the musical stress as a function of time which copes with a large variety of music signals. In addition, we use a multi-path dynamic programming algorithm to provide temporal stability as well as a robust multi-periodicity tracking, even in the presence of arrhythmic or slight musical passages. Compared to a previous variant of our algorithm [34], the major changes in this new version consist in incorporating a dynamic programming block and in avoiding any thresholding (neither hard or adaptive). These upgrades have notably increased the system performance and robustness. However, it appears that further effort should be devoted to the musical stress module to improve the overall system performance. In fact, a significant number of errors are the consequence of non-detected or over-detected attacks in the musical stress profile. This is especially the case for signals containing tenuous attacks or predominant vocal passages. Although the current system displays a high performance when computing the main tempo, future work is still needed to obtain a complete and structured metrical description of a musical piece that will fully exploit the information related to the metrical levels that is provided by the dynamic programming stage. If the reader is interested, a detailed list containing the name of excerpts used during the evaluation, the BPM annotations and all algorithm results can be found on-line at `www.tsi.enst.fr/~grichard/jasp06/`.

Acknowledgement

The authors would like to thank the anonymous reviewers for their constructive comments, suggestions and corrections.

References

[1] R. Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music Perception*, vol. 11, no. 4, 1994.

[2] D. Moelants, "Preferred tempo reconsidered," in *Proc. of the 7th Int. Conf. on Music Perception and Cognition*, 2002, pp. 580–583.

[3] F. Lerdahl and R. Jackendoff, *A generative theory of tonal music*, MIT Press, Cambridge, Massachusetts, 1983.

[4] T. Jehan, "Event-synchronous music analysis/synthesis," in *Proc. Int. Conf. on Digital Audio Effects*, 2004.

[5] C. Raphael, "Automatic segmentation of acoustic music signals using hidden Markov models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 4, pp. 360–370, april 1999.

[6] M. Goto and Y. Muraoka, "Real-time rhythm tracking for drumless audio signals," in *Proc. of the IJCAI*, 1997.

[7] O. Gillet and G. Richard, "Drum track extraction from polyphonic music signals," in *Proc. IEEE Workshop on App. Signal Proc. to Audio and Acoust. (WASPAA)*, 2005.

[8] A. Klapuri, A. Eronen, and J. Astola, "Automatic estimation of the meter of acoustic musical signals," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 1, 2006.

[9] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, 2002.

[10] P. Desain and H. Honing, "Computational models of beat induction: the rule based approach," *Journal of new music research*, vol. 11, no. 4, pp. 29–42, 1999.

[11] S. Hainsworth, *Techniques for the automated analysis of musical audio*, Ph.D. thesis, University of Cambridge, December 2003.

[12] M. Goto and Y. Muraoka, *Computational Auditory Scene Analysis*, chapter Music understanding at the beat level: real-time beat tracking for audio signals, Lawrence Erlbaum Associates, 1998.

[13] J. Seppänen, "Tatum grid analysis of musical signals," in *Proc.IEEE Workshop on App. Signal Proc. to Audio and Acoust. (WASPAA)*, 2001.

[14] F. Gouyon, P. Herrera, and P. Cano, "Pulse-dependent analyses of percussive music," in *Proc. of AES22 Int. Conf. on Virtual, Synthetic and Entertainment Audio*, 2002.

[15] K. Jensen and T. Andersen, "Beat estimation on t he beat," in *Proc.IEEE WASPAA*, 2003, pp. 87–90.

[16] A. P. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE ICASSP*, 1999.

[17] E. Scheirer, "Tempo and beat analysis of acoustic music signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, 1998.

[18] W. Sethares and T. Staley, "Meter and periodicity in musical performance," *J. of New Music Research*, vol. 30, no. 2, 2001.

[19] M. Alonso, R. Badeau, B. David, and G. Richard, "Musical tempo estimation using noise subspace projections," in *Proc. IEEE Workshop on App. Signal Proc. to Audio and Acoust. (WASPAA)*, 2003.

[20] S. Hainsworth and M. Macleod, "Beat tracking with particle filtering algorithms," in *Proc. IEEE WASPAA*, 2003, pp. 91–94.

[21] W. Sethares, R. Morris, and J. Sethares, "Beat tracking of musical performances using low-level audio features," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 2, 2005.

[22] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *Audio Eng. Soc.*, vol. 51, no. 4, 2004.

[23] J. Foote and S. Uchihashi, "The beat spectrum: a new approach to rhythm analysis," in *Proc. IEEE ICME*, 2001.

[24] M. Alonso, B. David, and G. Richard, "Tempo extraction for audio recordings," in *Proc. Mirex*, 2005, `http://www.music-ir.org/evaluation/ mirex-results/audio-tempo/index.html`.

[25] R. Badeau, R. Boyer, and B. David, "EDS parametric modeling and tracking of audio signals," in *Proc. of the 5th. Int. Conf. on DAFx*, 2002.

[26] R. Badeau, *Méthodes à haute résolution pour l'estimation et le suivi de sinusoïdes modulées. Application aux signaux de musique* (in French), Ph.D. thesis, Télécom Paris, France, April 2005.

[27] R. Badeau, B. David, and G. Richard, "Yet another subspace tracker," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[28] P. Vaidyanathan, *Multirate systems and filter banks*, Prentice-Hall PTR, 1992.

[29] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[30] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals in presence of white noise," *Journal of Multivariate Analysis*, vol. 20, no. 1, pp. 1–25, 1986.

[31] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 1, 2005.

[32] M. Alonso, G. Richard, and B. David, "Extracting note onsets from musical recordings," in *Proc. IEEE Int. Conf. on Multimedia & Expo (ICME)*, 2005.

[33] M. Dvornikov, "Formulae of numerical differentiation," math.NA/0306092, 2003.

[34] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of music signals," in *Proc. Int. Symposium on Music Inf. Retriev (ISMIR)*, 2004.

[35] R. Meddis, "Simulation of auditory-neural transduction:Further studies," *J. Acoust. Soc. Am.*, vol. 83, no. 3, pp. 1056–1063, March 1988.

[36] B. Moore, Ed., *Hearing*, Academic Press, 2nd edition, 1995.

[37] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall PTR, 1993.

[38] G. Peeters, "Time variable tempo detection and beat marking," in *Proc. of the ICMC*, 2005.

[39] F. Gouyon, "Quantitative comparison of tempo induction algorithms," `http://www.iua.upf.es/mtg/ismir2004/contest/tempoContest/node3.html`.

[40] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *Trans. on Speech and Audio Proc.*, vol. 14, no. 5, 2006.

[41] M. Goto and Y. Muraoka, "Issues in evaluating beat tracking systems," in *Proc. IJCAI-97 workshop on issues in AI and music*, 1997, pp. 9–16.

[42] D. Temperley, "An evaluation system for metrical models," *Computer Music Journal*, vol. 28, no. 3, pp. 28–44, 2004.

[43] Audio Tempo Extraction, "Music Information Retrieval Evaluation eXchange," 2005, `http://www.music-ir.org/evaluation/ mirex-results/audio-tempo/index.html`.

[44] X. Serra, *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*, Ph.D. thesis, Stanford University, USA, 1989.

[45] D. Schwartz, *Méthodes statistiques à l'usage des médecins et des biologistes*, Flammarion medecine series, third edition, 1963, In French.

[46] K. Hermus and P. Wambacq, "Assessement of signal subspace based speech enhancement for noise robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[47] J. F. Wang, C-H. Yang, and K-H. Chang, "Subspace tracking for speech enhancement in car noise environments," in *Proc. IEEE ICASSP*, 2004.