Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 67215, 14 pages
doi:10.1155/2007/67215

Research Article

Template-Based Estimation of Time-Varying Tempo

Geoffroy Peeters

IRCAM - Sound Analysis/Synthesis Team, CNRS - STMS, 1 pl. Igor Stravinsky, 75004 Paris, France

Received 1 December 2005; Revised 17 July 2006; Accepted 10 September 2006

Recommended by Masataka Goto

We present a novel approach to automatic estimation of tempo over time. This method aims at detecting tempo at the tactus level for percussive and nonpercussive audio. The front-end of our system is based on a proposed reassigned spectral energy flux for the detection of musical events. The dominant periodicities of this flux are estimated by a proposed combination of discrete Fourier transform and frequency-mapped autocorrelation function. The most likely meter, beat, and tatum over time are then estimated jointly using proposed meter/beat subdivision templates and a Viterbi decoding algorithm. The performances of our system have been evaluated on four different test sets among which three were used during the ISMIR 2004 tempo induction contest. The performances obtained are close to the best results of this contest.

1. INTRODUCTION

Tempo and beat are among the most important percepts of (western) music (a time structured set of sound events). Given the inherent ambiguity of tempo due to the various possible interpretations of the metrical structure of a rhythm, its automatic estimation remains a difficult task for a large variety of music genres. For this reason and given the number of potential applications, it is still the subject of an increasing number of research.

Western music notation represents musical events using a hierarchical metrical structure that distinguishes various time scales. For a typical three-level hierarchy, the smallest scale corresponds to the tatum period, the middle one to the tactus period, the largest one to the period of the musical measure. The *tatum* period can be defined as "the regular time division that mostly coincides with all note onsets" [1] or as the "shortest durational values in music that are still more than accidentally encountered" [2]. The *tactus* period is the perceptually most prominent period. It is the rate at which most people would tap their feet or clap their hands in time with the music. In many cases, this value corresponds to the denominator of the time signature [3]. In this paper, we deal with the estimation of the tempo at the tactus level, that is, the rate of the tactus pulse. It is expressed as number of beats per minute (BPM). The *musical measure* period corresponds to the description found in a score in the time signature and the bar lines. It is related to the harmonic change rate or to the length of a rhythmic pattern [2].

Many applications rely on tempo and beat information. Tempo can be used in search engines to query large databases and create automatically playlists based on tempo constraints. Some softwares or hardwares allow DJs to mix two tracks beat-synchronously or to synchronize sound devices with a given track. Audio sequencers based on the loop paradigm automatically extract the tempo and beat information to perform on-the-fly loop adaptations. (The loop paradigm consists in repeating (looping) many times a short extract of audio, such as a drum pattern, the length of which is chosen as an integer number of measures.) Recent creative paradigms use beat slicing (segmentation into beat units) as the base musical material. Music transcription and audio to score synchronization also benefit from the tempo and beat information. More generally, tempo can be considered as a periodicity reference for music such as pitch is for monophonic harmonic sounds. It can then be used for further audio analysis (beat-synchronous analysis).

However, many existing algorithms for automatic tempo and beat estimation make strong assumptions on the music content such as presence of periodical hard strikes (percussion/drum onsets), binary subdivision of the rhythm (usually a 4/4 meter is considered) or steadiness of the tempo over time. While these assumptions can be accepted for a large part of commercial music, it cannot be so when considering the whole diversity of (western) music including jazz, classical, and traditional music.

In this paper, we describe a system for the estimation of time-varying tempo and meter of a musical piece from the analysis of its audio signal. The system has been designed in order to allow this estimation for music with and without percussion. The front-end of the system is based on a *reassigned spectral energy flux* for the location of the musical events. A new periodicity measure based on a *combination of discrete Fourier transform and frequency-*

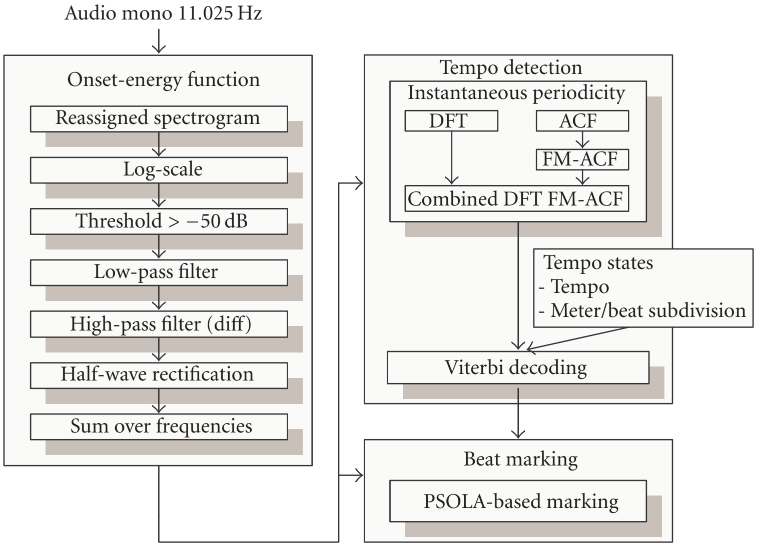


Figure 1: Flowchart of our system for tempo, meter estimation, and beat marking.

mapped auto-correlation function is proposed which allows a better discrimination between various existing periodicities (tatum, tactus, measure). A *Viterbi decoding* algorithm then estimates simultaneously the most likely tempo and meter over time using proposed meter/beat subdivision templates. The system is noncausal (therefore non real-time) since it uses information from future events (through the length of the analysis window and the use of a Viterbi algorithm). The flowchart of the system is represented in Figure 1.

Numerous studies exist concerning tempo and beat estimation. We refer the reader to [4] for a recent report on state-of-the-art tempo estimation algorithms. Using the taxonomy proposed in [4], we briefly review current directions in order to locate our algorithm in the field. Tempo estimation algorithms can first be distinguished from the analyzed materials: symbolic data [5, 6] or audio data. Algorithms based on audio analysis usually start by a front-end which either plays the role of an "audio-to-symbolic" translator (extract the exact location of the onsets of the events) [7–11] or extracts frame-based audio features such as energy, energy variations, energy in subbands or chord changes [2, 12, 13]. In the latter case, the features should represent significant cues concerning the presence of musical events and (or) their roles in the metrical structure. Depending on the kind of information provided by this front-end and the context of the application (real-time beat tracking or offline tempo estimation), a large variety of processes are used to track/estimate the tempo. In the case of a sequence of onsets, time interval histograms (inter-onset-histogram [8, 14]) are often used to detect the main periodicities. In the case of frame-based features, a periodicity measure (Fourier transform, autocorrelation function, narrowed-ACF [15], wavelets, comb filterbank) is mostly used. The periodicity measure can be used to estimate directly the tempo or to serve as observation for the estimation of the whole metrical structure through (probabilistic) models: estimation of the tatum, tactus (beat), measure and (or) estimation of systematic time deviations such as the swing factor [2, 11, 16, 17].

Paper organization

The paper is organized as follows. In Section 2, we present the front-end of our system for the extraction of the onset-energy function based on a proposed reassigned spectral energy flux. This onset-energy function is then used to estimate the dominant periodicities at each time. In Section 3.1, we present a new periodicity measure based on a combination of discrete Fourier transform and frequency-mapped autocorrelation function. In Section 3.2, we present our probabilistic model of tempo, the meter/beat subdivision templates and the Viterbi decoding algorithm which allows the estimation of the most likely tempo and meter path over time. In Section 4, we evaluate the performances of our system on four different test sets among which three were used during the ISMIR 2004 tempo induction contest.

2. ONSET-ENERGY FUNCTION

In order to detect the tempo of a piece of music from an audio signal, one needs first to extract meaningful information in terms of musical periodicity from the signal. This is the goal of the front-end of any audio-based tempo estimation algorithm. Front-ends can perform onset detection. However, by experimenting with this approach, we found it unreliable considering the consequences that false positive and false negative detections can have on the subsequent stages of the tempo estimation process. In [18] it has also been found that algorithms based on onset detection suffer more from distortion of the signal than the ones based on frame features.[1] In addition to that the concept of discrete onsets remains unclear for a large class of sounds such as slow attack, slow transition between notes without an attack phase and slow transition between chords such as played by

[1] Note however that [14] argues that a weak onset detector is suitable for tempo induction.

a string section. When front-ends extract frame-based audio features, the most commonly used features are the variation of the signal energy or its variation inside several frequency bands [12]. Since our interest is not only in music with percussion but also in music without percussion, our function should also react to any musically meaningful variations such as note transitions at constant global energy or slow attacks. These variations are usually visible in a spectrogram representation. Reference [17] proposes a function, called the spectral energy flux, which measures the variation of the spectrogram over time. For the computation of the spectrogram, [17] uses a window of length about 10 ms. This would lead according to [19] to a spectral resolution[2] of about 200 Hz. This spectral resolution is too large for the detection of transitions between adjacent notes especially in the lowest frequencies. In order to achieve such detection, one would need a much longer window, but then this would be to the detriment of the temporal precision of onset locations. This is the usual time versus frequency resolution trade-off. One would need a short window for accurate temporal location of percussive onset and a long window for accurate detection of transition between adjacent notes.

For this reason, we propose to compute the spectral energy flux using the reassigned spectrogram instead of the normal spectrogram. By using phase information, the reassigned spectrogram allows significant improvement of temporal and frequency resolution, therefore avoiding attacks blurring and better differentiation of very close pitches. Because of that, we argue that using a single long window with the reassigned spectrogram is suitable for onset detection for both percussive and nonpercussive audio.

2.1. *Reassigned spectrogram*

In the following, we call "bin" a specific point of the short time Fourier transform grid defined by its frequency ω_k and time t_m. The reassigned spectrogram [20] consists of reallocating the energy of the "bins" of the spectrogram to the frequency ω_r and time t_r corresponding to their center of gravity. It has already been used for applications such as transient detection, glottal closure instant detection in speech, sinusoidality coefficient or harmonic frequency location [21–24]. The reassignment of the frequencies is based on the computation of the instantaneous frequency which is the time derivative of the phase. We note x the signal, h the analysis window of length L centered on time t_m, dh the time derivative of the window $h(dh = \partial h(t)/\partial t)$, STFT_h the short time Fourier transform computed using h, and STFT_{dh} the one computed using dh. The reassignment of the frequencies can be efficiently computed by

$$\omega_r(x, t_m, \omega_k) = \omega_k - \Im\left\{\frac{\text{STFT}_{dh}(x, t_m, \omega_k)}{\text{STFT}_h(x, t_m, \omega_k)}\right\}, \qquad (1)$$

where $\Im$ stands for the imaginary part. The reassignment of the times is based on the computation of the group delay which is the frequency derivative of the phase spectrum. We note th the frequency derivative of the window $h(th = t \cdot h(t))$ and STFT_{th} the short time Fourier transform computed using th. The reassignment of the times can be efficiently computed by

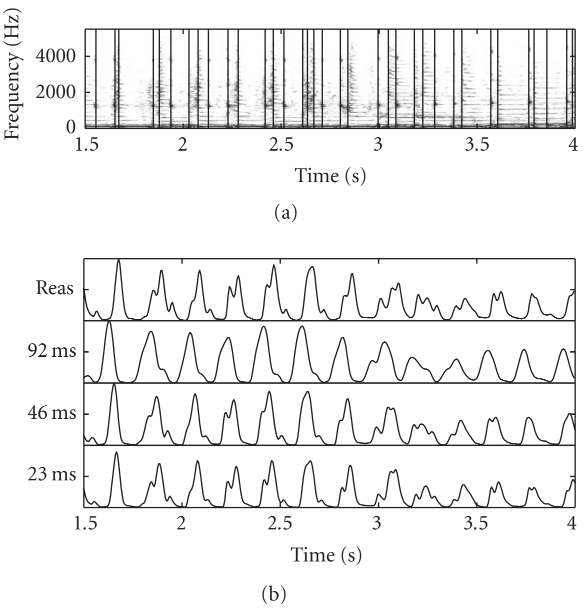


FIGURE 2: From top to bottom: (a) reassigned spectrogram computed using a window length of 92.8 ms, superimposed: manually annotated onset locations, (b1) corresponding reassigned spectral energy flux function, (b2) normal spectral energy flux function computed using a window length of 92 ms, (b3) 46 ms, (b4) 23 ms on [signal: Asian Dub Foundation, RAFI, track 01 "Assassin" from the "songs" database of the ISMIR 2004 test set].

$$t_r(x, t_m, \omega_k) = t_m + R\left\{\frac{\text{STFT}_{th}(x, t_m, \omega_k)}{\text{STFT}_h(x; t_m, \omega_k)}\right\}, \qquad (2)$$

where R stands for the real part.

Each "bin" (ω_k, t_m) of the spectrogram is then reassigned to its center of gravity (ω_r, t_r) using (1) and (2). Since ω_r and t_r are real-valued, we round them to the closest discrete frequency $\omega_{k'}$ and discrete time $t_{m'}$ of the STFT grid. The bins are finally accumulated in the time and frequency plane.

2.2. *Reassigned spectral energy flux*

Except for the use of reassigned spectrogram, the computation of the reassigned spectral energy flux is close to the computation of the normal spectral energy flux. It is done in the following way.

(1) The signal is first down-sampled to 11.025 Hz and converted to mono (mixing both channels).

(2) The reassigned spectrogram $X(\omega_{k'}, t_{m'})$ is computed using a hamming window. A long window of 92.8 ms (1023 samples) is used in order to achieve a good frequency resolution. This favors the detection of note changes in the spectrum and therefore high values in the spectral flux. The decrease of the time resolution due to the use of a long window is compensated by the use of the group delay (see Figure 2

[2] For two sinusoidal components of equal amplitude, the spectral resolution is the minimal distance between their frequencies that guarantee that no overlap between their main lobe occurs above a −3 dB level. The spectral resolution depends on the window length and shape.

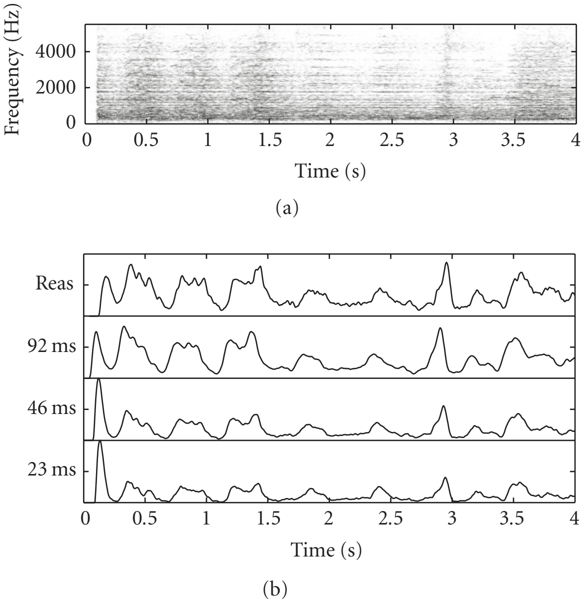


Figure 3: Same as Figure 2 but on [signal: Bernstein conducts Stravinsky, track 23 "The jovial merchant with two gypsy girls" from the "songs" database of the ISMIR 2004 test set].

and the corresponding discussion below). The number of bins of the DFT used in (1) and (2) is 1024. The hop size is set to 5.8 ms (64 samples).

(3) As in [7], the energy spectrum is converted to the log scale. The use of the log scale will allow us in step (4) to work on variations of energy relative to the energy level since $\partial \log(A(t))/\partial t = (\partial A(t)/\partial t)/A(t)$. A threshold of 50 dB below the maximum energy is applied.

(4) The energy inside each frequency band $e_{\log}(\omega_k, t_m)$ is low-pass filtered with an elliptic filter of order 5 and a cutoff frequency of 10 Hz. The goal of the low-pass filter is to avoid the detection of spurious onsets due to the presence of background noise or noise events such as cymbal sounds. The resulting energy signals are then differentiated using a simple $[1, -1]$ differentiator. The number of frequency bands is among half the size of the DFT used in step (2), 500 in our case.

(5) The resulting energy signals $e_{\text{filter}}(\omega_k, t_m)$ are then half-wave rectified. We note them $e_{\text{HWR}}(\omega_k, t_m)$.

(6) For a specific time t_m, the sum over all frequency bands ω_k is computed: $e(t_m) = \sum_k e_{\text{HWR}}(\omega_k, t_m)$. The resulting energy function $e(n = t_m)$ has a sampling rate of 172 Hz.[3]

2.3. Comparison with the spectral energy flux

In Figures 2 and 3, we compare the reassigned and the normal spectral energy flux functions. The latter has been obtained by using the normal spectrogram instead of the reassigned spectrogram in step (2) of Section 2.2. Each figure represents the reassigned spectrogram using a window of length 92.8 ms, the corresponding reassigned spectral energy flux function, noted $e_{\text{reas}}(n)$, and three versions of the normal spectral energy flux functions computed using three different window lengths for the spectrogram (92.8 ms, 46.3 ms and 23.1 ms), noted $e_{92}(n)$, $e_{46}(n)$, and $e_{23}(n)$, respectively. Figure 2 represents the results for percussive audio (rock music) and Figure 3 for nonpercussive audio (classical music). In the case of percussive audio, we have superimposed the manual annotation of the onset locations to the reassigned spectrogram. In Figure 2, it can be seen that many of the percussive onsets visible in $e_{\text{reas}}(n)$ are missing in $e_{92}(n)$. This comes from the blurring that occurs on the normal spectrogram due to the use of a long window. In this case, a shorter window is needed in order to highlight the onsets in $e(n)$ as the one used for $e_{23}(n)$. In Figure 3, we observe the inverse behavior. Many onsets visible in $e_{\text{reas}}(n)$ are missing in $e_{23}(n)$. This comes from the weak frequency resolution obtained using a short window. In this case, a longer window is needed in order to highlight the onsets in $e(n)$, as the one used for $e_{92}(n)$. In the case of the spectrogram, both types of signal would thus require a different window length. We see that with a single window length, the reassigned spectrogram succeeded to highlight the onsets in both cases.

We continue this comparison in Section 4.3.1 where we evaluate the influence of the choice of the reassigned or normal spectral energy flux function as well as the influence of the window length on the global tempo recognition rate.

3. TEMPO DETECTION

We estimate the tempo from the analysis of the onset-energy function $e(n)$. The algorithm we propose works in two stages: (i) first we estimate the dominant periodicities at each time (Section 3.1); (ii) then we estimate the tempo, meter, and beat subdivision paths that best explain the observed periodicities over time (Section 3.2).

3.1. Periodicity estimation

Periodicity estimation of a signal is often done using discrete Fourier transform (DFT) or autocorrelation function (ACF). Ideally, $e(n)$ is a periodic signal that can be roughly modeled as a pulse train convolved with a low-pass envelope. If we note $f = f_0$ for fundamental frequency, the outcome of its DFT is a set of harmonically related frequencies $f_h = hf_0$. Depending on their relative amplitude it can be difficult to decide which harmonic corresponds to the tempo frequency. If we note $\tau = 1/f_0$ the period of $e(n)$, the outcome of its ACF is a set of periodically related lags $\tau_h = h/f_0$. Here also it can be difficult to decide which period corresponds to the tempo lag. Algorithms like the two-way mismatch [8, 25] or maximum likelihood [26] try to solve this problem. In [27] we have proposed a more straightforward approach that we apply here to the problem of tempo periodicity estimation.

3.1.1. Combined DFT and frequency-mapped ACF

The octave uncertainties of the DFT and ACF occur in inverse domains: frequency domain $f_h = hf_0$ for the DFT, lag domain $\tau_h = h/f_0$, or inverse frequency domain $f_h = f_0/h$ for the ACF. We use this property to construct a combined

[3] Note that one could easily derive the onset locations by applying a threshold on $e(n)$.

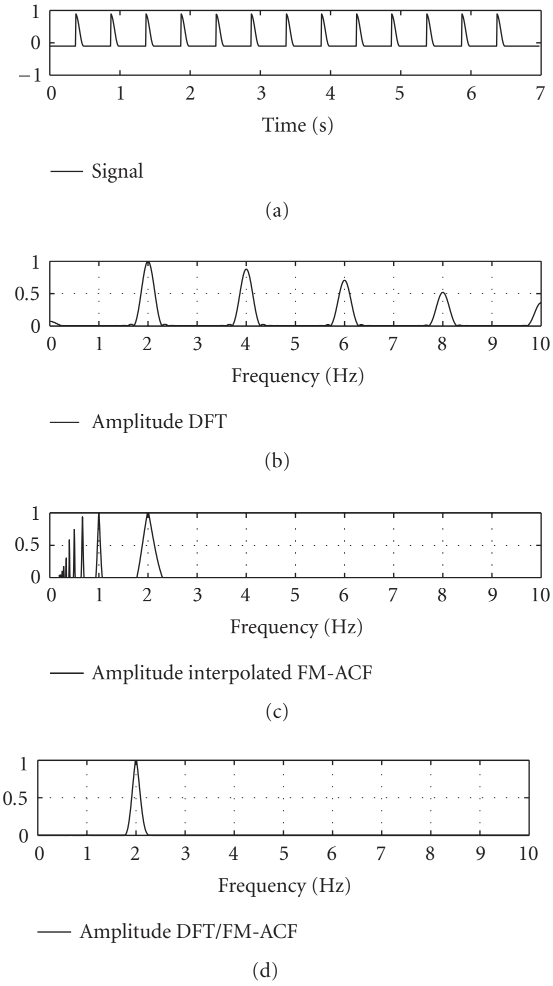


FIGURE 4: Simple example of combination between the DFT and the ACF. From top to bottom: (a) signal, (b) magnitude of the DFT, (c) ACF function mapped to the frequency domain, (d) product of (b) and (c); on [signal: periodic impulse signal at 2 Hz].

function that reduces these uncertainties. We believe this combined function can be very useful for the detection of the various periodicities of a rhythm since it allows to better discriminate the various periodicities of the measure, tactus, and tatum (see Figure 6 in the remaining).

Example 1. In Figure 4, we illustrate the principle of the method with a simple example. Figure 4(a) represents a periodic impulse signal at 2 Hz, Figure 4(b) its DFT, Figure 4(c) its ACF mapped to the frequency domain (the lags τ_l are represented as frequencies $f_l = 1/\tau_l$), Figure 4(d) the product of the DFT and this frequency-mapped ACF. Only the component at $f = f_0$ remains.[4]

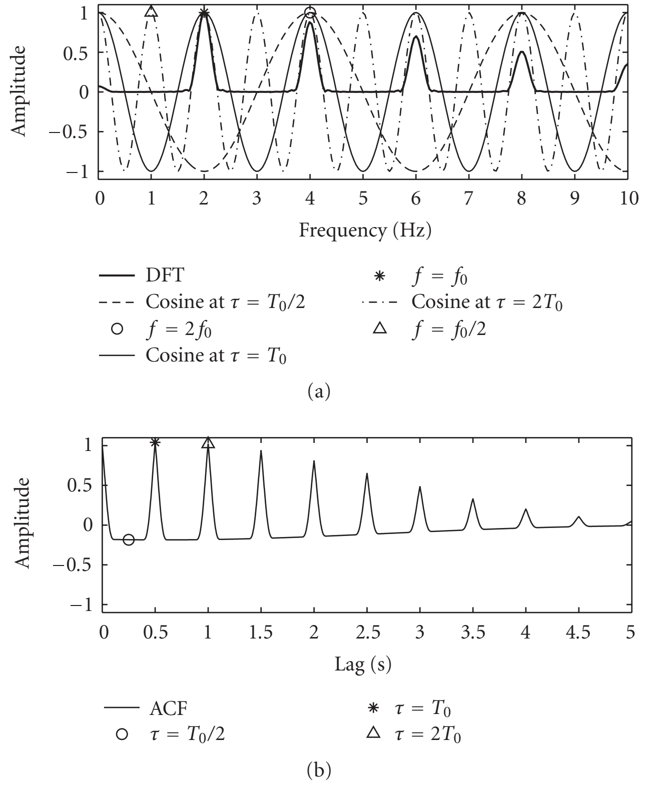


FIGURE 5: (a) magnitude of the DFT of the signal; superimposed: cosine at $\tau = T_0/2$, T_0, $2T_0$ and $f = 2f_0$, f_0, $f_0/2$ positions; (b) autocorrelation function; superimposed: $\tau = T_0/2$, T_0, $2T_0$ positions; on [signal: periodic impulse signal at 2 Hz].

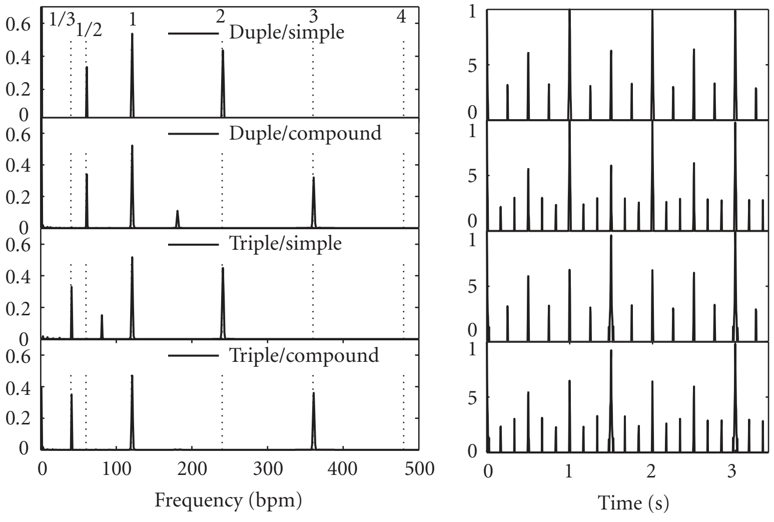
Explanations

This interesting property comes from the fact that the ACF $\hat{r}(\tau)$ of a signal is equal to the inverse Fourier transform of its power spectrum $|S(\omega)|^2$. Since the power spectrum is real and symmetric, its (inverse) Fourier transform reduces to the real part. Therefore, $\hat{r}(\tau)$ can be considered as the projection of $|S(\omega)|^2$ on a set of cosine functions $g_\tau(\omega) = \cos(\omega\tau)$ with frequencies equal to the lag τ. In other words, $\hat{r}(\tau)$ measures the periodicity of the peak positions of the power spectrum.

Example 2. In Figure 5, we illustrate this for a periodic impulse signal at $f_0 = 2$ Hz. We decompose $g_\tau(\omega)$ into its positive and negative parts: $g_\tau(\omega) = g_\tau^+(\omega) - g_\tau^-(\omega)$. Positive values of $\hat{r}(\tau)$ occur only when the contribution of the projection of $|S(\omega)|^2$ on $g_\tau^+(\omega)$ is greater than the one on $g_\tau^-(\omega)$ (this is the case for the subharmonics of f_0, $\tau = k/f_0$, $k \in \mathbb{N}^+$ in the figure); nonpositive values when the contribution of $g_\tau^-(\omega)$ is larger than or equal to the one of $g_\tau^+(\omega)$ (this is the case for the higher harmonics of f_0, $\tau = 1/(kf_0)$, $k > 1$, $k \in \mathbb{N}^+$ in the figure). It is easy to see that only for the value $\tau = 1/f_0$ we have simultaneously a maximum of the projection of $|S(\omega)|^2$ on $g_\tau(\omega)$ and a peak of energy in $|S(\omega)|^2$ at $f = 1/\tau$.

This inverse octave uncertainty of the DFT and ACF is used to compute our new periodicity measure as follows.

[4] In this example, we rely on the fact that energy exists in the DFT at the frequency $f = f_0$. In order to solve a possible "missing fundamental" (no energy at $f = f_0$), we have proposed in [27] the use of the autocorrelation of the DFT instead of the use of the direct DFT. In this paper, we will however use the direct DFT.



(a)

(b)

Figure 6: (a) Metrical patterns of the combined DFT/FM-ACF for a tempo of 120 bpm and various theoretical typical rhythms; (b) corresponding temporal signals.

Computation

We first make $e(n)$ a zero-mean unit-variance signal. $e(n)$ is then analyzed both by the following.

(1) *DFT*: we note $S(\omega_k, t_m)$ the magnitude spectrum of $e(n)$ for a frequency ω_k and a frame centered around time t_m. A hamming window is used with length equal to 8 s. The hop size is set to 0.5 s.

(2) *Frequency mapped ACF (FM-ACF)*: we note $\hat{r}(\tau_l, t_m)$ the autocorrelation function of $e(n)$ for a lag τ_l and a frame centered around time t_m. This function is normalized in length and in maximum value. The normalized-in-length autocorrelation function is defined as

$$r(l,m) = \frac{1}{L-l}\sum_{n=0}^{L-l-1} e\left(n+m-\frac{L}{2}\right)e\left(n+l+m-\frac{L}{2}\right), \quad (3)$$

where l is the lag τ_l expressed in samples, m the time of the frame t_m in samples, and L the window length in samples. The normalization in maximum value (at the zeroth-lag) is obtained by $\hat{r}(l) = r(l)/r(0)$. A rectangular window is used with length equal to 8 s. The hop size is set to 0.5 s.

The value $\hat{r}(\tau_l, t_m)$ represents the amount of periodicity of the signal at the lag τ_l or at the frequency $\omega_l = (2\pi)/\tau_l$ for all $l > 0$. Each lag τ_l is therefore "mapped" in the frequency domain. Of course since $\hat{r}(\tau_l, t_m)$ has a constant resolution in lag, $\hat{r}(\omega_l, t_m)$ has a decreasing resolution in frequency. In order to get the same linearly spaced frequencies ω_k as for the DFT, we interpolate[5] $\hat{r}(\tau_l, t_m)$ and sample it at the lags $\tau'_l = (2\pi)/\omega_k$. For this computation, we only consider the frequencies ω_k corresponding to tempo values between 30 and 600 bpm ($\omega_k \in [0.5, 10]$ Hz, $\tau'_l \in [0.1, 2]$ s). Finally, half-wave rectification is applied to $\hat{r}(\omega_k, t_m)$ in order to consider only positive auto-correlation.

(3) *Combined function*: the DFT and the FM-ACF provide two measures of periodicity at the same frequencies ω_k. We finally compute a combined function $Y(\omega_k, t_m)$ by multiplying the DFT and the FM-ACF at each frequency ω_k:

$$Y(\omega_k, t_m) = S(\omega_k, t_m) \cdot \hat{r}(\omega_k, t_m). \quad (4)$$

In the following $Y(\omega_k, t_m)$ will be considered as our signal observation.

Choice of a window length

The length of the window used for the computation of the DFT and the ACF affects the interpretation one can make concerning the observed periodicities. Short windows tend to capture tatum periodicity, middle ones tactus periodicity, and long ones periodicity of the measure. For a 120 bpm musical piece, the length of a beat period is 0.5 s. In order to discriminate the beat frequencies in a spectrum (to avoid spectral leakage), one would need a length larger than 2 s (4 time the period length). Also, in order to observe the periodicity of the measure this would lead to 8 s for a 4/4 meter, our choice for the system. We also apply a zero-padding factor of 4.[6] The number of frequencies ω_k of the DFT is therefore equal to 8192 bins[7] and the distance between two frequencies is equal to 1.26 bpm (0, 021 Hz). The hop size is set to 0.5 s.

In the left part of Figure 6, we represent the patterns of $Y(\omega_k)$ for various theoretical typical rhythm characteristics

[5] Note that this does not improve the frequency resolution of $\hat{r}$.

[6] The number of bins of the DFT is taken as 4 times the smallest power of two that is greater than or equal to the window length.

[7] Note however that we only consider the frequencies corresponding to tempo values between 30 and 600 bpm.

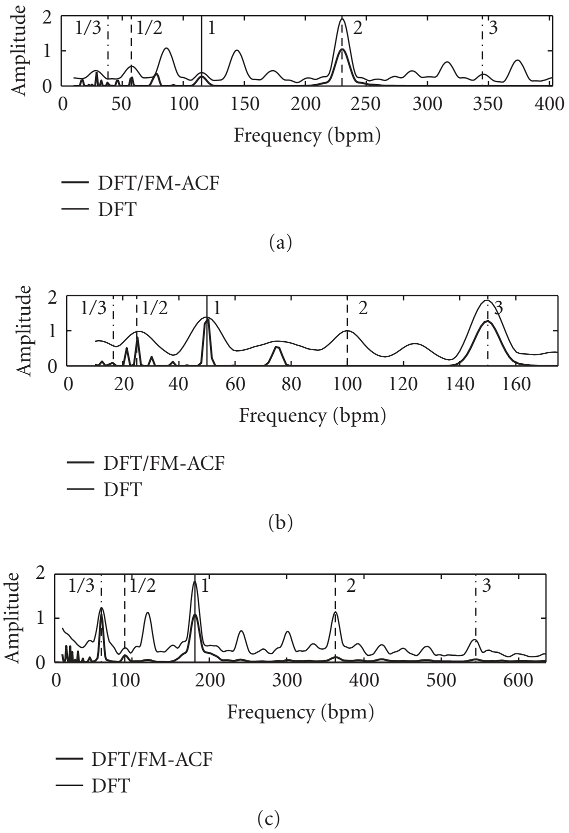


Figure 7: Comparison between the DFT (thin line) and the combined DFT/FM-ACF (thick line) measured on real signals: (a) quadruple/simple meter, (b) duple/compound meter, (c) triple/simple meter. Superimposed: ground-truth tempo (1), 1/2 and 2 time the tempo, 1/3 and 3 time the tempo.

and a tempo of 120 bpm: duple/simple meter (eighth note at 2/4), duple/compound meter (6/8), triple/simple meter (eighth note at 3/4), triple/compound meter (9/8). In the upper part of the figure the integer number 1 refers to the tactus, the highest peak to the right (2 or 3) is the tatum and the highest peak to the left (1/2 or 1/3) to the measure level. The resulting patterns of $Y(\omega_k)$ are simple. This comes from the fact that $Y(\omega_k)$ is the product of two inverse periodic series based on the periodicity of the measure (kf_m) and of the tatum (f_t/k'). Figure 6(b) represents the corresponding temporal signal. The tactus period is equal to 0.5 s.

In Figure 7, we compare the mean values over time of $S(\omega_k, t_m)$ and $Y(\omega_k, t_m)$, noted $\overline{S}(\omega_k)$ and $\overline{Y}(\omega_k)$, measured on real signals. The signal represented in Figure 7(a) is a quadruple/simple meter.[8] Remark the large difference between the values taken by $\overline{S}(\omega_k)$ and $\overline{Y}(\omega_k)$. The value at the tempo frequency (1) is much more emphasized in $\overline{Y}(\omega_k)$ than in $\overline{S}(\omega_k)$. Figure 7(b) represents a duple/compound meter.[9] As in Figure 6, we observe the typical 1, 3 pattern in $\overline{Y}(\omega_k)$. Figure 7(c) represents a triple/simple meter.[10] As in Figure 6, we observe the typical 1/3, 1 pattern in $\overline{Y}(\omega_k)$. In all these cases, $\overline{Y}(\omega_k)$ gives a better emphasis on the tempo and rhythm specificities than $\overline{S}(\omega_k)$.

3.2. Tempo estimation

The dominant periodicities $Y(\omega_k, t_m)$ are estimated at each time t_m. As depicted in Figure 6, $Y(\omega_k, t_m)$ does not only depend on the tempo (120 bpm in Figure 6) but also on the characteristics of the rhythm, at least on the subdivision of the meter and of the beat. We therefore look for the temporal path of tempo and meter/beat subdivision that best explains $Y(\omega_k, t_m)$.

Tempo states

In the following we consider three different kinds of meter/beat subdivisions, named meter/beat subdivision templates (MBST):

(i) the duple/simple (noted 22 in the following),
(ii) the duple/compound (noted 23, example is 6/8 meter) and
(iii) the triple/simple (noted 32, example is 3/4 meter).

We define a "tempo state" as a specific combination of a tempo frequency b_i and an MBST $m_j : s_{ij} = [b_i, m_j]$ with $i \in I$ the set of considered tempo and $j \in \{22, 23, 32\}$ the three considered MBSTs. We look for the most likely temporal succession of "tempo states" given our observations. We formulate this problem as a Viterbi decoding algorithm [28].[11]

Viterbi decoding algorithm

Viterbi decoding algorithm, as used in HMM decoding [29], requires the definition of three probabilities: an emission probability of the states $p_{\text{emi}}(Y(\omega_k, t_m) \mid s_{ij}(t_m))$, a transition probability between two states $p_t(s_{ij}(t_{m+1}), s_{kl}(t_m))$, and a prior probability of each state $p_{\text{prior}}(s_{ij}(t_0))$.

The *emission probability* $p_{\text{emi}}(Y(\omega_k, t_m) \mid s_{ij}(t_m))$ is the probability that the model emits a given signal observation $Y(\omega_k, t_m)$ at time t_m given that the model is in state s_{ij} at time t_m. This probability could be learned from annotated data as we did in [30].[12] In the present system, we use a more straightforward computation based on the theoretical metrical patterns represented in Figure 6. For a specific tempo b_i and MBST m_j, we first compute a score defined as a weighted

[8] Enya, Watermark, "Orinoco flow," [Rhino/Warner Bros].

[9] Boyz II Men, Coolexhighharmony, "End of the road" [Motown].

[10] Viennese Waltz "media104409" from the "ballroom-dancer" database of the ISMIR 2004 test set.

[11] Our method shares some similarities with [17] in the use of a dynamic programming technique. Reference [17] uses it to estimate simultaneously the most likely tempo and downbeat location over time based on the observation of the energy flux signal and considering only a duple/simple meter. We use it here to estimate simultaneously the most likely tempo and meter/beat subdivision over time based on the observation of $Y(\omega_k, t_m)$.

[12] It should be noted that in [31] a weighted sum of specific ACF periodicities has also been proposed in a task of meter and tempo estimation.

sum of the values of $Y(\omega_k, t_m)$ at specific frequencies:

$$\text{score}_{i,j}(Y(\omega_k, t_m)) = \sum_{r=1}^{5} \alpha_{j,r} \cdot Y(\omega = \beta_r \cdot b_i, t_m), \tag{5}$$

where $\underline{\beta}$ represents the various ratios of the considered frequency ω to the tempo frequency b_i of the state s_{ij},

$$\underline{\beta} = \left[\frac{1}{3}, \frac{1}{2}, 1, 1.5, 2, 3\right]. \tag{6}$$

These ratios correspond to significant frequency components for the triple meter, duple meter, tempo, "penalty" (see below), simple and compound meter. $\underline{\alpha}_j$ represents the weightings of each of these components. These weightings depend on the MBST m_j of the state s_{ij} and have been chosen to better discriminate the various MBSTs:

$$\begin{aligned} \underline{\alpha}_{22} &= [-1, 1, 1, -1, 1, -1] \quad \text{if } m_j = 22, \\ \underline{\alpha}_{23} &= [-1, 1, 1, -1, -1, 1] \quad \text{if } m_j = 23, \\ \underline{\alpha}_{32} &= [1, -1, 1, -1, 1, -1] \quad \text{if } m_j = 32. \end{aligned} \tag{7}$$

The ratio $\beta = 1.5$ is called the "penalty" ratio. It is used to reduce the confusion between 22 and 23/32 MBST. Indeed, the eighth note frequency of a rhythm at x bpm in a 22 MBST (tactus at the quarter note) can be interpreted as the eighth note triplet frequency of a rhythm at $(2/3)x$ bpm in a 23 MBST (tactus at the dotted quarter note).[13] The negative weighting given to the ratio 1.5 penalizes these choices.

The probability that state s_{ij} emits a given signal observation is based on this score and is computed as

$$p_{\text{emi}}(Y(\omega_k, t_m) \mid s_{ij}(t_m)) = \frac{\text{score}_{i,j}(Y(\omega_k, t_m))}{\sum_{i,j} \text{score}_{i,j}(Y(\omega_k, t_m))}. \tag{8}$$

The *transition probability* favors continuity of tempi and MBST over time. We consider independence between tempo and MBST.[14] We compute this probability as the product of a tempo continuity probability and an MBST continuity probability,

$$\begin{aligned} p_t&(s_{ij}(t_{m+1}) \mid s_{kl}(t_m)) \\ &= p_t(b_i(t_{m+1}) \mid b_k(t_m)) \cdot p_t(m_j(t_{m+1}) \mid m_l(t_m)). \end{aligned} \tag{9}$$

The goal of the first probability is to favor continuous tempi. We set it as a Gaussian pdf $N_{\mu=b_k,\sigma=5}(b_i)$. The goal of the second probability is to avoid MBST jumps from frame to frame. We set it empirically to 0.0833 for $j \neq l$ and 0.833 for $j = l$.

The *prior probability* $p_{\text{prior}}(s_{ij}(t_0))$ is the prior probability to observe a specific tempo i and a specific MBST j. This probability is set according to musical knowledge. Assumptions about tempo range and meter can be made according to the music genre of the track. This music genre could be automatically estimated by including a front-end for music genre recognition in our system. Since our current system does not include such a front-end, we simply favor the detection of tempo in the range 50–150 bpm but we do not favor any MBST in particular. We set it as a Gaussian pdf: $p_{\text{prior}}(s_{ij}(t_0)) = p_{\text{prior}}(b_i(t_0)) = N_{\mu=120,\sigma=80}(b_i)$.

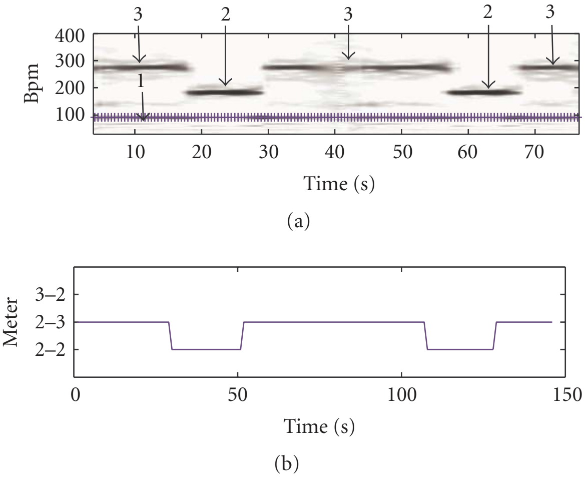


Figure 8: (a) tempo estimation over time (b) MBST estimation over time; on [signal: "Standard of excellence-accompaniment CD-Book2-All inst.-88. Looby Loo"].

A standard Viterbi decoding algorithm is then used to find the best path of states $[b_i, m_j]$ over time, which gives us simultaneously the best tempo and MBST path that explain $Y(\omega_k, t_m)$. Finally, in order to increase the precision of the tempo estimation, frequency interpolation is performed around the value $Y(b(t_m), t_m)$. For this a second-order polynomial, $p(\omega) = a\omega^2 + b\omega + c$, is fitted to the values of $Y(\omega_k, t_m)$ around $\omega_k = b(t_m)$. The value corresponding to the maximum of the polynomial, $\omega_{\max} = -b/(2a)$, is chosen as the final tempo value.

Example 3. In Figure 8 we illustrate the estimation of time-varying MBST. Figure 8(a) represents the estimated tempo track over time (indicated with "+"s around 100 bpm) superimposed to the periodicity observation $Y(\omega_k, t_m)$ represented as a matrix and annotated by hand (1 for tactus frequency, 2 and 3 for tatum frequency). Figure 8(b) represents the estimated MBST over time. The system has estimated a constant tempo during the entire track duration but depending on the local periodicities (1 and 3 or 1 and 2), the MBST is estimated as either 23 or 22. Both tempo and MBST estimations are correct.

Example 4. In Figure 9, we illustrate the estimation of time-varying tempo on Brahms "Ungarische Tanze n5."[15] This

[13] The same is true for the sixteenth note and a rhythm at $(4/3)x$ bpm in a 23 MBST.

[14] This is not exactly true since some joint tempo/meter transitions are more likely than others.

[15] The track has been annotated by hand into beat locations. The local tempo has then been derived from the distance between adjacent beats. Note that the resulting tempo would not necessarily correspond to the perceived tempo.

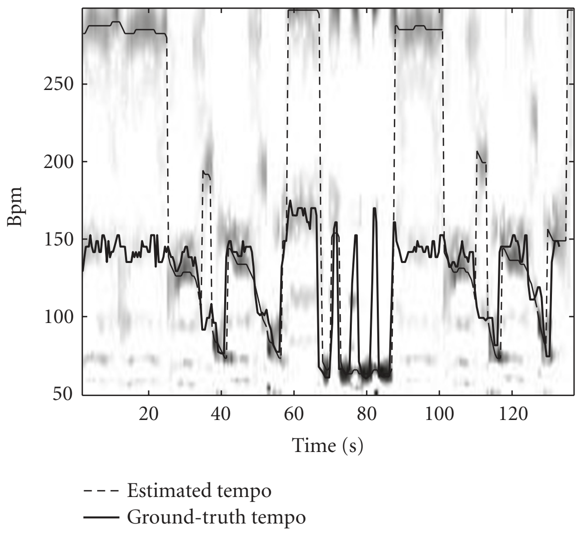


Figure 9: Tempo estimation over time: estimated tempo (dashed line), ground-truth tempo (continuous thick line) on [signal: Brahms "Ungarische Tanze n5"].

piece is interesting since it has many quick tempo variations. The dashed thin line represents the estimated tempo track while the continuous thick line represents the reference tempo. Both are superimposed to the observations matrix $Y(\omega_k, t_m)$. The tempo has been estimated as twice the reference tempo during the periods $[0, 25]$, $[34, 37]$, $[58, 67]$, $[88, 101]$, and $[110, 113]$ s and as half during the period $[75, 85]$ s. The transitions being very quick in this part, the algorithm decided there was a higher probability to remain at 65 bpm.

4. EVALUATION

In this section, we evaluate the performances of our tempo estimation system.

4.1. Test sets

Evaluation of algorithms is often done on personal test sets. However, this makes the comparison with existing technologies hard. For this reason, and because of availability, we used the three test sets of the ISMIR 2004 tempo induction contest (see [18] for details). We also added a fourth "personal" test set in order to represent also commercial radio music. The test sets are

(i) the "ballroom-dancer" database:[16] 698 tracks of 30 s long. The following music genres are covered: cha cha, jive, quickstep, rumba, samba, tango, Viennese waltz and slow waltz music. The tracks are mainly in 4/4 and 3/4 meters and with almost constant tempo except for the slow waltz music,

(ii) the "songs" database: 465 tracks of 20 s long. The following music genres are covered: rock, classical, electronica, latin, samba, jazz, afrobeat, flamenco, Balkan and Greek music. The tracks are in various meters and with constant or time variable tempo (flamenco, classical),

Table 1: Comparison between reassigned and normal spectral energy flux for various window lengths in a task of tempo estimation.

	11.5 ms		23, 1 ms		46, 3 ms		92, 8 ms	
	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2
RSEF	48, 0	79, 4	49, 5	82, 4	49, 9	83, 2	49, 5	83,7
SEF	49, 7	80, 4	49, 5	82, 6	49, 3	82, 8	49, 7	82, 2

(iii) the "loops" database: 1889 tracks of "loops" to be used in DJ sessions from the Tape Gallery.[17] Although the database used in [18] had 2036 items, we had only access to 1889 of them (92.8%). Also we had to manually correct part of the annotations since some of them did not represent any musical meaningful periodicities. When comparing our results with the ISMIR 2004 results, one should keep that in mind. It is also worth to mention that, despite of its name, the database contains a large part of non drum-loops sounds like machine/engine noises with unclear periodicity,

(iv) the "poprock" database: 153 tracks of 20 s covering commercial radio music from the last decades (80's, 90's, 00's, including pop, rock, rap, musical comedy).

In the following, the results obtained with our system will be compared with the ones obtained during the ISMIR 2004 tempo induction contest published in [18]. Each item of the four test sets has been annotated by its mean tempo over time. The "ballroom-dancer" and "poprock" databases have also been annotated by the author in meter. We have used the three following meters: 22 (if the annotated beats can be musically grouped by 2 and subdivided by 2), 23 (grouped by 2 divided by 3), 32 (grouped by 3 divided by 2).

The tracklist of the "poprock" database, as well as the used tempo and meter annotations for the four test sets can be found on the author's web site.[18]

4.2. Evaluation method

The tempo over time was extracted with our algorithm. The tempo was not considered constant during the track duration. For each track, we compare the median value of the estimated tempo over time with the annotated tempo. As in [18], we consider two accuracy measures:

(i) accuracy 1: percentage of tempo estimates within 4% of the ground-truth tempo,

(ii) accuracy 2: percentage of tempo estimates within 4% of either the ground-truth tempo, 1/2, 2, 1/3 or 3 the ground-truth tempo. This allows taking into account the fact that various periodic levels often coexist within a given metric. Because the ground-truth meter is available for the "ballroom-dancer" and "poprock" databases, we also indicate a more restrictive definition of accuracy 2 that only considers the estimated tempo as correct when it is 1/2, 1 or 2 for the 22 meter, 1/3, 1 or 2 for 32 meter, 1/2, 1 or 3 for 23 meter.

[16] http://www.ballroomdancers.com.

[17] http://www.sound-effects-library.com.

[18] http://recherche.ircam.fr/equipes/analyse-synthese/ peeters/eurasipbeat/.

Table 2: Results of the tempo estimation evaluation.

	Ballroom		Songs		Loops		Poprock	
	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2	Acc1	Acc2
Time variable 22/23/32	65,2	93,1 (89,0)	49,5	83,7	56,1	80,7	87,6	97,4 (97,4)
Constant 22	68,7	96,9	39,4	85,2	59,8	83,1	81,7	99,4
ISMIR 2004 best	63,2	92,0	58,5	91,2	70,7	81,9		

4.3. Results

4.3.1. Comparison between reassigned and normal spectral energy flux

We first compare the results obtained using various choices for the front-end of our system. We test the choice of the reassigned or normal spectral energy flux, noted RSEF and SEF, respectively. In both cases, we test the influence of the window length, noted L. Four lengths are tested: $L = 11.5$ ms, 23.1 ms, 46.3 ms, and 92.2 ms. For this comparison, we only use the "songs" database since this is the most balanced database among the four, containing both percussive and nonpercussive audio. In Table 1, we indicate the accuracies 1 and 2 of the whole system for the eight versions of the front-end. According to accuracy 1, all choices lead to close results except for the choice of the RSEF with $L = 11.5$ ms which has the lowest score. According to accuracy 2, the RSEF with $L = 92.8$ ms slightly outperforms the other methods.[19] This therefore confirms the choice we have made previously. It is interesting to consider that also for $L = 46.3$ ms, the RSEF slightly outperforms the SEF. For both RSEF and SEF, the lowest score is obtained with $L = 11.5$ ms, the choice made in [17].

The results presented in the following are obtained with the reassigned spectral energy flux and a window of length 92.6 ms.

4.3.2. Evaluation of the system

In Table 2, we compare the results obtained using our system ("time variable 22/23/32" row) with the best results obtained during the ISMIR 2004 tempo induction contest ("ISMIR 2004 best" row). We indicate the accuracies 1 and 2 for the four test sets. The values in parentheses correspond to the restrictive accuracy 2.

In Figures 10, 11, 12, and 13 we present detailed results for each database. We define r as the ratio between the estimated tempo and the ground truth tempo. The upper part of each figure (a) represent the histogram of the values r in log-scale over all instances of each database. The vertical lines represent the values of r corresponding to usual tempo confusions: 1/3, 1/2, 2/3, 4/3, 2, 3 ($-1.58, -1, -0.58, 0.41, 1, 1.58$ in log-scale). The lower part of each figure (b) indicates the influence of the precision window width on the recognition rate. The vertical line represents the precision window width of 4% used in Table 2.

For the "*ballroom-dancer*" database, the results are 65.2%/93.1% (89.0) which improve upon those obtained in ISMIR 2004 (63.2%/92.0%). Considering accuracy 1, most errors occurred in the jive and quickstep (half the tempo), rumba (twice the tempo) and both waltzes. The jive and quickstep explains the large peak at $r = 1/2$ in the histogram of Figure 10. Considering accuracy 2, most errors occurred in the slow waltz (the concept of onsets is unclear in the slow chord transitions). We also evaluate the recognition rate of the ground-truth meter. Comparing the estimated meter with the ground-truth meter makes sense only for track with correctly estimated tempo.[20] The recognition rate of meter (for the 65.2% remaining tracks) is 88.7% for the 22 meter (3.8% recognized as 23, 7.4% as 32), 43.9% for the 32 meter (51.6% recognized as 22, 4.4% as 23). This is surprisingly low.

For the "*songs*" database, the results are 49.5%/83.7% which is lower than those obtained in ISMIR 2004 (58.5%/91.2%) but would be the second best algorithm according to accuracy 2. The large difference between accuracies 1 and 2 (and the high peak in the histogram of Figure 11 at $r = 2$) indicates that in many cases the algorithm estimated the tatum periodicity. Despite our 1.5 penalty coefficient, a secondary peak exists in the histogram at $r = 2/3$ (detection of the dotted quarter note). According to Figure 11, increasing the width of the precision window to more than 4% would increase a lot accuracy 2.

For the "*loops*" database, the results are 56.1%/80.7%, just below those obtained in ISMIR 2004 (70.7%/81.9%) but would be the second/third best algorithm. Three peaks exist in the histogram at $r = 0.5$, $r = 2$, and $r = 4/3$.

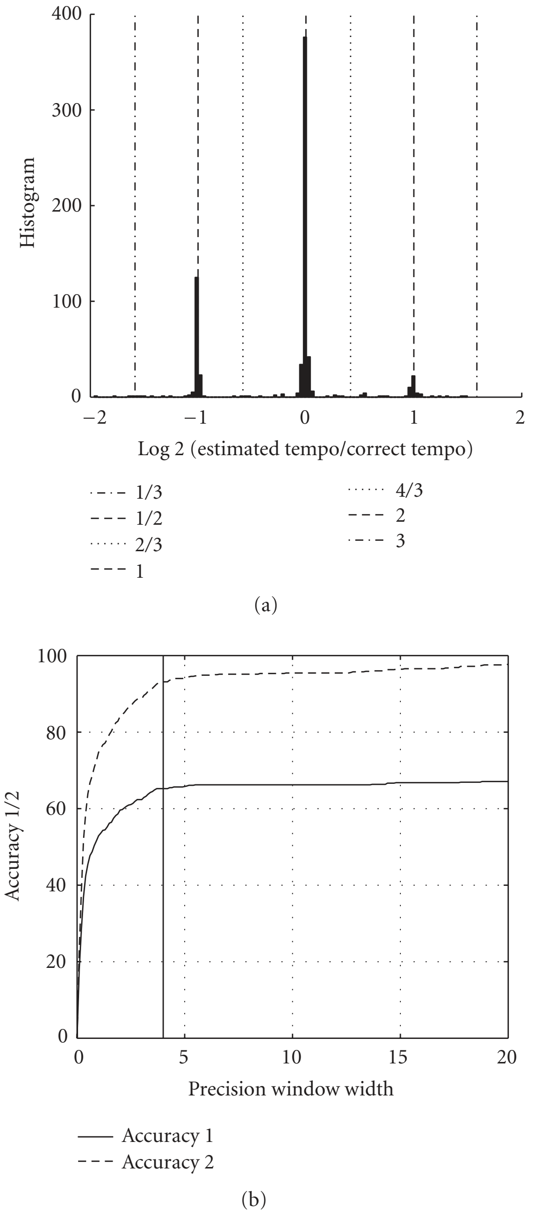
For the "*poprock*" database, the results are 87.6%/97.4% (97.4%). The recognition rate of meter (for the 87.6% correctly estimated tempo) is 89.3% for the 22 meter (3% recognized as 23, 7.6% as 32), 100% for the 23 meter.

In order to check the importance of the meter/beat subdivision and the time-varying estimation (Viterbi decoding) parts of our algorithm, we have done the evaluation again with a constant tempo and a 22 meter/beat subdivision hypothesis. For this, we only estimate the most likely $p_{\text{emi}}(\overline{Y(\omega_k)} \mid [b_i, 22])$ of (8) and only using an average observation over time $\overline{Y(\omega_k)}$. In this case, the weightings of (7) are defined as $\underline{\alpha} = [0, 1, 1, 0, 1, 0]$, that is, we did not use any penalty weightings. The results are indicated in Table 2 ("Constant 22" row).

Surprisingly, for the *ballroom-dancer* database, both accuracies increase by about 3.5%. In this case, the evaluation

[19] Since the database contains 465 titles, a difference of 0.21% indicates a difference of one correct recognition.

[20] A track with a 32 meter will not be estimated as 32 if the estimated tempo is twice the ground-truth tempo.

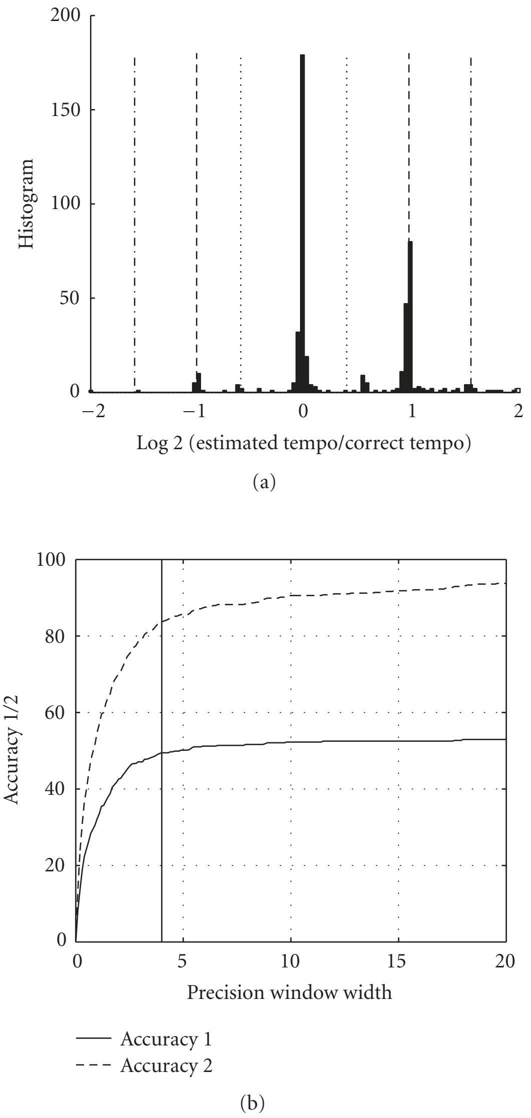


(a)

(b)

FIGURE 10: (a) Histogram of the ratios in log-scale between estimated tempi and correct tempi; (b) accuracy versus precision window width (in (%) of correct tempo) for the ballroom-dancer database.

of MBST has a negative effect on the result. For the *songs* database, accuracy 1 decreases by almost 10% while accuracy 2 increases by 1.5%. The evaluation of MBST has therefore a positive impact on accuracy 1, that is, it allows avoiding confusion between the various levels of the metrical structure. For the *loops* database, both accuracies increase by about 3%. This is normal since the given hypothesis (constant tempo and duple/simple meter) is largely valid for this database. It is interesting to note that the simplified algorithm now outperforms in accuracy 2 (83.1%) the best results of ISMIR 2004 (81.9%). For the *poprock* database, accuracy 1 decreases by 6% while accuracy 2 increases by 2%. Here also, the evaluation of MBST has a positive impact on accuracy 1.



(a)

(b)

FIGURE 11: Same as Figure 10 for the songs database.

As a conclusion, when given no prior knowledge about tempo evolution over time and meter/beat subdivision, the use of the proposed MBST increases accuracy 1 (except for the ballroom-dancer) and slightly decreases accuracy 2. When constant tempo and duple/simple meter hypothesis holds, the use of MBST has a negative effect.

CONCLUSION AND DISCUSSIONS

The system presented in this paper yields very good performance for tempo estimation for a large variety of music genres. Among the three test sets used for the ISMIR 2004 tempo induction contest, our system outperformed once the previous best results and was close to them for the two others. However, the automatic estimation of the meter, based on the proposed meter/beat subdivision templates, remains unreliable.

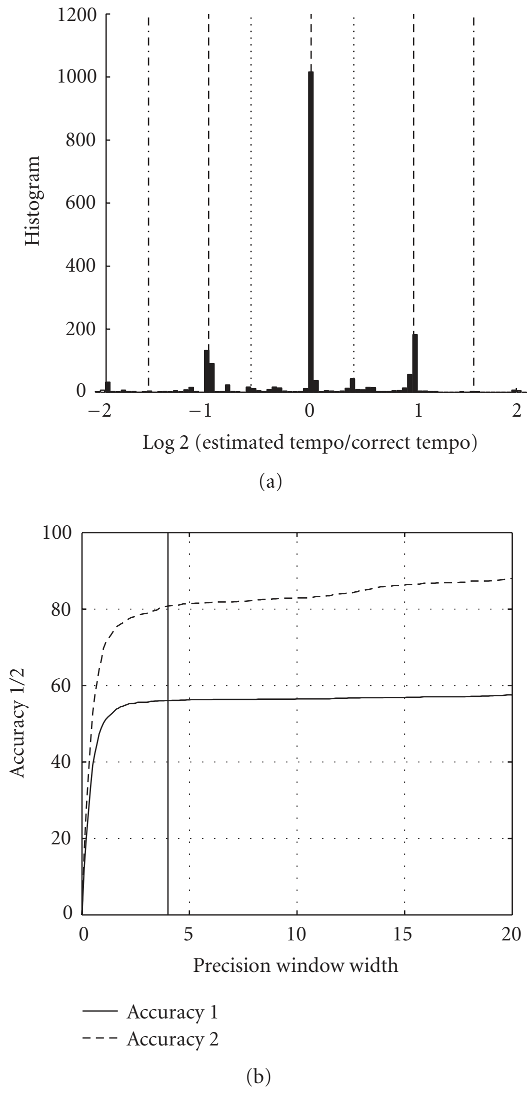


Figure 12: Same as Figure 10 for the loops database.

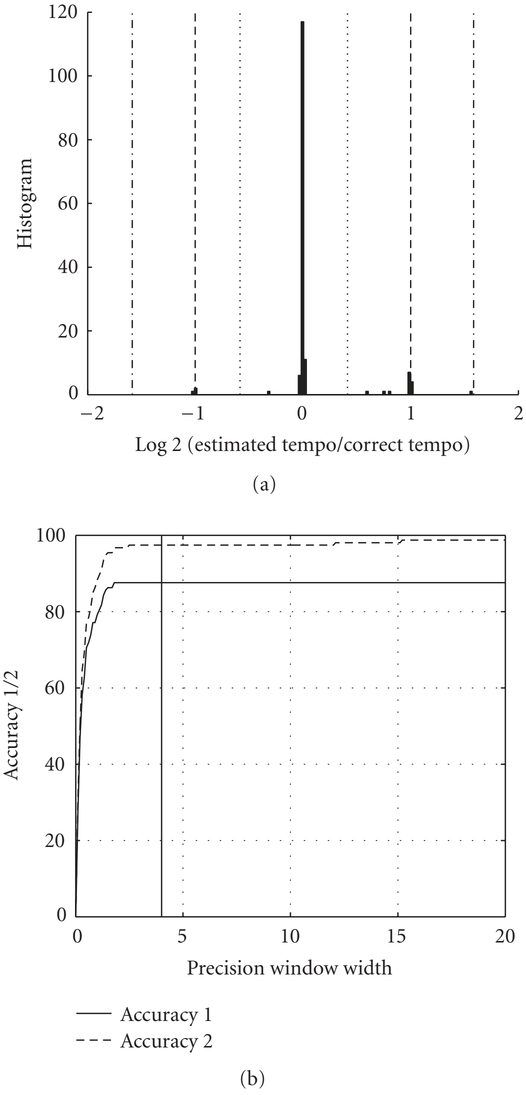


Figure 13: Same as Figure 10 for the poprock database.

Trying to improve our system, we should distinguish two main problems. The first one concerns the extraction of significant information from the audio signal that allows the estimation of a musical periodicity. For this, we have shown in an experiment that the proposed *reassigned spectral energy flux* using a long analysis window can provide slight improvement over the usual spectral energy flux especially for nonpercussive audio. We also base this assertion on the first place obtained by our system in the nonpercussive audio category of the MIREX 2005 tempo contest.[21] However, the sole information extracted from the signal is related to energy (energy variations). This information is surely too poor for the characterization of rhythm [17]. Inclusion of features such as pitch, relative frequency positions, spectral centroid/spread [3] could certainly improve the performances of our system.

The second problem concerns the estimation of the tempo itself. Because the tempo has inherent ambiguities due to the various possible interpretations of a metrical structure of a rhythm, we have proposed to estimate it jointly with the measure and tatum periodicities through the use of *meter/beat subdivision templates*. This was possible since the proposed *combined DFT/ FM-ACF* allows a good discrimination between the measure, tactus, and tatum periodicities. Considering the performance of the tempo estimation, we believe this approach is promising. However, considering the performance of the estimated meters, there is space for improvements. There are two reasons for that. The first reason comes from the weighting used in the templates that are based on theoretical templates. These templates only represent the variety of possible existing rhythm patterns partially. One solution would be to learn the templates from annotated

[21] http://www.music-ir.org/mirex2005/index.php/Audio_Tempo_Extraction for details.

data as we did in [30]. In the current work, we did not want to use this information from the test sets. The second reason comes from signal processing. The interpolation used during the mapping of the ACF to the frequency domain degrades the resolution of the combined function in the low frequencies (where the measure/bar frequency is located). The meter subdivision estimation is therefore more difficult than the beat subdivision estimation. Among the most problematic rhythms (except those in exotic meters) are the ones with accentuations on dotted quarter notes that are frequent in bossa-nova or funk music. Specific templates should be devoted to that as well.

As represented in Figure 1, the system also contains a beat marking algorithm which we did not discuss here since it was not possible to evaluate because of the lack of annotated databases for beat locations. For the same reason, the time-varying characteristics of our algorithm have only been indirectly tested in the median-tempo evaluation. Ongoing work will concentrate on these improvements and evaluations.

ACKNOWLEDGMENTS

Part of this work was conducted in the context of the European IST project Semantic HIFI[22] [32]. The author would like to thank three anonymous reviewers for useful and detailed suggestions. Many thanks also to the people who have collaborated to the ISMIR 2004 tempo induction contest and made the test sets and annotations available. To my father.

REFERENCES

[1] J. Bilmes, “Timing is of the essence: perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm,” M.S. thesis, MIT, Cambridge, Mass, USA, 1993.

[2] A. Klapuri, A. Eronen, and J. Astola, “Analysis of the meter of acoustical musical signals,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[3] F. Gouyon, *A computational approach to rhythm description*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005.

[4] F. Gouyon and S. Dixon, “A review of automatic rhythm description systems,” *Computer Music Journal*, vol. 29, no. 1, pp. 34–54, 2005.

[5] J. C. Brown, “Determination of the meter of musical scores by autocorrelation,” *Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 1953–1957, 1993.

[6] P. Allen and R. Dannenberg, “Tracking musical beats in real time,” in *Proceedings of the International Computer Music Conference and International Computer Music Association*, pp. 140–143, San Francisco, Calif, USA, September 1990.

[7] A. Klapuri, “Sound onset detection by applying psychoacoustic knowledge,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’99)*, vol. 6, pp. 3089–3092, Phoenix, Ariz, USA, March 1999.

[8] F. Gouyon, P. Herrera, and P. Cano, “Pulse-dependent analyses of percussive music,” in *Proceedings of AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pp. 396–401, Espoo, Finland, June 2002.

[9] J. Bello, *Towards the automated analysis of simple polyphonic music: a knowledge based approach*, Ph.D. thesis, Queen Mary University of London, London, UK, 2003.

[10] C. Uhle and J. Herre, “Estimation of tempo, micro time and time signature from percussive music,” in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx ’03)*, pp. 84–89, London, UK, September 2003.

[11] M. Goto, “An audio-based real-time beat tracking system for music with or without drum-sounds,” *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[12] E. D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[13] J. Paulus and A. Klapuri, “Measuring the similarity of rhythmic patterns,” in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR ’02)*, pp. 150–156, Paris, France, October 2002.

[14] S. Dixon, “Automatic extraction of tempo and beat from expressive performances,” *Journal of New Music Research*, vol. 30, no. 1, pp. 39–58, 2001.

[15] J. C. Brown and M. S. Puckette, “Calculation of a “narrowed” autocorrelation function,” *Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1595–1601, 1989.

[16] F. Gouyon and P. Herrera, “Determination of the meter of musical audio signals: seeking recurrences in beat segment descriptors,” in *Proceedings of the 114th Convention of Audio Engineering Society (AES ’03)*, Amsterdam, The Netherlands, March 2003.

[17] J. Laroche, “Efficient tempo and beat tracking in audio recordings,” *Journal of the Audio Engineering Society*, vol. 51, no. 4, pp. 226–233, 2003.

[18] F. Gouyon, A. Klapuri, S. Dixon, et al., “An experimental comparison of audio tempo induction algorithms,” *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, pp. 1832–1844, 2006.

[19] F. J. Harris, “On the use of windows for harmonic analysis with the discrete Fourier transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[20] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, Academic Press, San Diego, Calif, USA, 1999.

[21] G. Peeters and X. Rodet, “Sinola: a new analysis/synthesis using spectrum peak shape distortion, phase and reassigned spectrum,” in *Proceedings of the International Computer Music Conference (ICMC ’99)*, pp. 153–156, Beijing, China, October 1999.

[22] G. Peeters, *Modèles et modélisation du signal sonore adaptés à ses caractéristiques locales*, Ph.D. thesis, Université Paris VI, Paris, France, 2001.

[23] A. Röbel, “A new approach to transient processing in the phase vocoder,” in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx ’03)*, pp. 344–349, London, UK, September 2003.

[24] S. Hainsworth and P. Wolfe, “Time-frequency reassignment for music analysis,” in *Proceedings of International Computer Music Conference (ICMC ’01)*, pp. 14–17, La Habana, Cuba, September 2001.

[25] R. Maher and J. Beauchamp, “Fundamental frequency estimation of musical signals using a two-way mismatch procedure,” *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.

[26] B. Doval and X. Rodet, “Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’93)*, vol. 1, pp. 221–224, Minneapolis, Minn, USA, April 1993.

[27] G. Peeters, “Music pitch representation by periodicity measures based on combined temporal and spectral representations,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’06)*, pp. 53–56, Toulouse, France, May 2006.

[22] http://shf.ircam.fr.

[28] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[29] L. R. Rabiner, “Tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[30] G. Peeters, “Rhythm classification using spectral rhythm patterns,” in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR ’05)*, pp. 644–647, London, UK, September 2005.

[31] S. Dixon, E. Pampalk, and G. Widmer, “Classification of dance music by periodicity patterns.,” in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR ’03)*, pp. 159–165, Baltimore, Md, USA, October 2003.

[32] H. Vinet, “The Semantic Hifi project,” in *Proceedings of the International Computer Music Conference (ICMC ’05)*, pp. 503–506, Barcelona, Spain, September 2005.

Geoffroy Peeters was born in Leuven, Belgium, in 1971. He received his M.S. degree in electrical engineering from the Université-Catholique of Louvain-la-Neuve, Belgium, in 1995 and his Ph.D. degree in computer science from the Université Paris VI, France, in 2001. During his Ph.D., he developed new signal processing algorithms for speech and audio processing. Since 1999, he works at IRCAM (Institute of Research and Coordination in Acoustic and Music) in Paris, France. His current research interests are in signal processing and pattern matching applied to audio and music indexing. He has developed new algorithms for timbre description, sound classification, audio identification, rhythm description, automatic music structure discovery, and audio summary. He owns several patents in these fields and received the ICMC Best Paper Award in 2003. He has also coordinated indexing research activities for the Cuidad, Cuidado, and Semantic HIFI European projects. He is the coauthor of the ISO MPEG-7 audio standard.