

Spectral and Temporal Periodicity Representations of Rhythm for the Automatic Classification of Music Audio Signal

Geoffroy Peeters

Abstract—In this paper, we study the spectral and temporal periodicity representations that can be used to describe the characteristics of the rhythm of a music audio signal. A continuous-valued energy-function representing the onset positions over time is first extracted from the audio signal. From this function we compute at each time a vector which represents the characteristics of the local rhythm. Four feature sets are studied for this vector. They are derived from the amplitude of the Discrete Fourier Transform, the Auto-Correlation Function, the product of the DFT and the ACF interpolated on a hybrid lag/frequency axis and the concatenated DFT and ACF coefficients. Then the vectors are sampled at some specific frequencies, which represent various ratios of the local tempo. The ability of these periodicity representations to describe the rhythm characteristics of an audio item is evaluated through a classification task. In this, we test the use of the periodicity representations alone, combined with tempo information and combined with a proposed set of rhythm features. The evaluation is performed using annotated and estimated tempo. We show that using such simple periodicity representations allows achieving high recognition rates at least comparable to previously published results.

Index Terms—Rhythm description, rhythm classification, automatic indexing, audio features.

I. INTRODUCTION

AUTOMATIC music description from signal analysis has become one of the important research fields in the last decade. Music description is often achieved by combining three different points of view [1]: melody/harmony, timbre (which is related roughly to the orchestration of the music), and tempo/rhythm. This last point raises questions about the *representation of time* into a compact and generalizable form that is suitable for task such as classification, search by similarity or visualization.

A. Motivating applications

The goal of this paper is to study the ability of four spectral and temporal periodicity representations to describe the rhythm content of a music audio signal (not a symbolic representation of music). The motivating applications are the development of systems for automatic classification of an audio signal into classes of rhythm, or the improvement of the performances of automatic classification of an audio signal into music genre and mood (using also the complementary melody/harmony and timbre features). A final motivating application is the inclusion of rhythm features in search-by-similarity algorithms usually only based on timbre features.

G. Peeters is with the Sound Analysis/Synthesis Team of Ircam - CNRS STMS, 75004 Paris, France (e-mail: geoffroy.peeters@ircam.fr).

B. Related works

For the representation of the rhythm content of an audio signal, several proposals have been made so far. The main differences between them are - the type of information being represented (representation of event positions, of the timbre characteristics of them or both) - the way they are represented (sequence of events, transform in the lag or frequency domain, histogram, sum over a similarity matrix or features derived from the previous), - the algorithm used to compare the representations (simple Euclidean or cosine distances, dynamic programming or elaborate machine learning algorithms). Rather than a chronological overview of related works, we propose to group them into five main approaches.

Similarity matrix based approach. Foote [2] proposes the use of a *beat spectrum* to visualize the temporal structure of a song (beat, measure and small structure). The beat histogram is obtained by computing the Self Similarity Matrix (SSM) of a signal (represented by its STFT amplitude coefficients or by its MFCCs, using either Euclidean or cosine distance) and - summing the values along diagonals at specific lags or - computing its auto-correlation. Various distances are proposed by Foote in [3] for comparing the beat spectrum of two tracks: Euclidean, cosine distance or the cosine distance between the Fourier coefficients of the beat spectrum. The SSM approach is also followed by Antonopoulos in [4]. For a given music track, the “chroma-based MFCC” features are extracted either from the whole signal or from an estimated thumbnail. From this, the SSM is computed and used to create a “rhythmic signature”. Dynamic Time Warping is then used to compute the distance between two signatures. Evaluation is performed on Greek Traditional Dance and African music.

Features based approach. Tzanetakis [1] proposes the use of a *beat histogram* obtained by collecting over time the contribution of the dominant peaks of an enhanced autocorrelation. Various features are derived from this histogram and used, in combination with timbre and pitch content features, for music genre classification. Paulus [5] proposes to model the rhythm characteristics as a sequence of audio features (for example loudness, brightness or MFCCs) over time. A Dynamic Time Warping algorithm is then used to align the time axis of two sequences and allow their comparison. Gouyon [6] proposes a set of 73 features to characterize the rhythm. Those include features derived from the tempo, from the *Periodicity Histogram* (Pampalk [7]) and from an *Inter-Onset-Interval Histogram* (IOIH). These features are used for the

classification of 8 music genres from the “Ballroom dancer” test-set. The authors report 90.1% correct recognition using the correct tempo (78.9% using the estimated tempo). Another study made by Gouyon [8] considers tempo estimation errors as part of the estimation process. They use 28 pair-wise classifiers and obtain 67.6% correct recognition.

Temporal pattern based approach. Dixon [9] proposes to add to the Gouyon set of features, a *temporal rhythmic patterns* derived from the energy evolution of the signal inside each bar. Various other features are also used (for example meter, syncopation and swing factor). On the “Ballroom dancer” test-set, the authors report 50% correct recognition using only the pattern, and up to 96% using this pattern and all features with an AdaBoost classifier. The results of 85.7% obtained in [9] using only rhythm pattern, Periodicity and IOI histogram (no tempo or bar length attributes) is often considered as the state of the art for rhythm classification without annotated tempo. However, it seems that the position of the first bars estimated using BeatRoot algorithm have been “corrected manually” (part 3.3 of [9]). Wright [10] also uses temporal-templates in the case of tempo, beat and downbeat estimation of Afro-Cuban music. Using a matched-filtering approach, they first enhance the presence of claves in the audio signal. The positions of the discrete onsets of the enhanced signal are then compared to a set of temporal-templates representing the theoretical positions of claves at various tempi and pattern rotations. A rotation-aware dynamic programming algorithm is then used to find the tempo, beat and downbeat positions.

Normalized/ scaled/ aligned periodicity measures approach. Holzapfel [11] proposes the use of Dynamic Periodicity Warping (DPW) to compute rhythmic similarity. An average-over-frame DFT of an onset-strength signal is first computed. The similarity between two tracks is computed using the cost of the DPW alignment of the bins of their respective DFTs. The authors study the use of the direct cost, a normalized cost (normalization by the cost of a reference path representing tempo variation), a cosine distance between the two aligned spectra and a direct Euclidean or cosine distance between the non-aligned spectra. On the “Ballroom dancer” test-set, the best accuracy (85.5%) is obtained using a weighted K-NN and the cosine distance between non-aligned spectra. It should be noted that this representation is dependent on the tempo (tempo information is intrinsically represented by the frequency localization of the DFT peaks). The best tempo-independent representation is the proposed normalized DPW with 82.1% accuracy. Another evaluation is performed on the “Six dances from Crete” test-set. Another study made by Holzapfel [12] uses the Melin Transform (MT) to provide a theoretically scale (and therefore tempo) independent rhythm representation. For this, the Fast Melin Transform (FMT) of the sample auto-correlation of an onset-function is computed. Since this representation is supposed to be tempo independent, a direct cosine distance between the FMT (using 40 coefficients) of two tracks is used to compute their similarity. On the “Ballroom dancer” test-set, the best accuracy (86.9%) is again obtained using a weighted K-NN and the cosine distance between the Periodicity Spectra (which are tempo-dependent). The best tempo-independent representation

is the proposed FMT with 85.1% accuracy (the previously proposed DPW reaches 84.0% accuracy). For the “Crete” test-set, the best approach is the tempo-independent FTM (77.4%). Jensen [13] proposes to exponentially group the lags of the auto-correlation function of an onset function. Each track is represented by the values of 60 exponentially spaced bands representing the lags between 0.1s and 4s. A modified Nearest Neighbors algorithm is used (looking also at slight shifts of the vectors to take into account small tempo variations) to perform classification. On the “Ballroom dancer” test-set, the proposed logarithmic-ACF method achieves 85.7% accuracy, while the usual linear ACF achieves 89% accuracy. It is worth to mention that, while independent of small tempo changes, the proposed logarithmic ACF representation is depend on tempo for changes larger than 3%. It therefore models intrinsically the tempo and cannot be said to tempo independent as can be seen in Figure 4 of [13]. Gruhne’s approach [14] is based on the beat histogram of Tzanetakis [1] (auto-correlation of the onset function). They propose the use of a logarithmically-spaced lag-axis in order to get rid of tempo changes. In order to compute it, they propose an algorithm for the estimation of a reference point. Results on two private test-sets show improvements over the usual linear-lag beat histogram for task of classification and similarity.

Source separation based approach. Other approaches rely on the instrument transcription of the percussive part of a song. For example, Uhle [15] uses Independent Subspace Analysis to transcribe the percussive part of a song and then estimate a histogram representing the frequency of occurrence of the percussive events on distinct metric positions. Tsunoo [16] proposes a method to describe rhythm using classification of track histograms. Ono’s [17] method is first used to extract percussive components from audio. Percussive patterns are then clustered using a combination of one-pass DP and k-means clustering algorithm. The frame of each track is then assigned to a cluster. The corresponding track’s histogram is used to perform classification using linear SVM. Results on the “Ballroom dancer” test-set indicate accuracies up to 69.1%.

The approach studied in this paper belongs to the Normalized/ scaled/ aligned periodicity measures approaches.

C. Paper content and organization

In this paper, we study the ability of four periodicity representations, in the spectral or temporal domain, to describe the rhythm content of a music audio signal. The paper is organized as follows.

In Section II, we present the four periodicity representations under consideration. We first set a list of requirements for a robust representation (II-A). We then review the four periodicity representations: the DFT (Section II-B), the ACF (Section II-C), a proposed Hybrid-Axis DFT-ACF (Section II-D) and a proposed concatenated DFT and ACF (Section II-E). We then discuss if these representations fulfill our requirements (Section II-F). We explain how these representations can be made tempo independent (Section II-F2) and compact (Section II-F3) through a sampling process. In Section III, we propose a small set of complementary rhythm features. In Section IV,

we compare the use of these four representations for a task of music genre classification. We study the use of the periodicity representations alone (computed using annotated or estimated tempo), when combined with tempo information (annotated or estimated) and when combined with the proposed small set of rhythm features. Using automatic feature selection, we then highlight the most discriminative features for our task of classification (Section IV-E). We conclude in Section V and give directions for future works.

D. Estimation of Tempo, meter and onset-energy function

In this work, we study various spectral and temporal representations. The periodicity representations are applied to a continuous-valued onset-energy-function (onset-strength signal) $o(n)$, where n represents the sample number. In this study we have used the reassigned-spectral-energy-flux function described in [18]. As explained in [18] this function facilitates the highlight of onsets for music with non-percussive instruments (such as for the Slow-Waltz class of our test-set). This function has a sampling rate of 172 Hz. We believe that the approach presented in this paper can be applied to any other continuous-valued onset-energy-function.

When sampling the periodicity axis (frequency or lag) of the representations, we use a reference tempo. We will study the use of annotated tempo but also the use of estimated tempo. The estimated tempo is the one provided by the algorithm proposed in [18]. When using annotated tempo, because the annotations only provide a global tempo and because part of the tracks of the test-set have time-variable tempo, the exact frequency of the local tempo is refined at each frame t_i using a comb-filter approach applied to the DFT coefficients. We denote this tempo by $f_{bpm}(t_i)$.

II. SPECTRAL AND TEMPORAL PERIODICITY REPRESENTATIONS

A. Introduction

Rhythm can be roughly defined by the tempo, the position and duration of the events and their timbre characteristics¹. Instead of the position of the events, we usually prefer to work on the sequence of event's relative position (or the successive Inter-Onset-Intervals). In this paper, we do not consider the timbre characteristics of the events but only the sequence of their relative positions. We seek a representation of the rhythm which fulfills the following requirements:

- 1) sensitivity to the sequence (order) of relative positions (but robust to small changes),
- 2) independence of the tempo (i.e. the speed of reading of the sequence) and
- 3) compactness.

We first review the four types of periodicity representations that will be used in this study. We then discuss if these representations fulfill our requirements.

For all frame-analysis, we use a window length of 8s and a hop size of 0.5s. For a signal at 120 bpm, this duration will

¹For example, Paulus [5] represents the characteristics of the events by their relative loudness, brightness, MFCCs, ...

allow the observation of 4 bars of 4 beats each. It therefore allows to obtain a good spectral resolution (no-overlapping at -3 dB of the main lobes of the DFT) for the harmonics of the bar frequencies. For a 30s length track, the total number of frames is 45 frames.

B. Amplitude of the Discrete Fourier Transform (DFT)

For each frame centered on t_i , we compute the amplitude of the DFT of the onset-energy-function $o(n)$ using a Hamming window. We denote it by $A(f_k, t_i)$ where f_k represents the frequencies. In order to have values of $A(f_k, t_i)$ independent of the local signal amplitude, $A(f_k, t_i)$ is normalized by its maximum amplitude. In the following we name it "DFT".

C. Auto-Correlation Function (ACF)

For each frame centered on t_i , we compute the Auto-Correlation Function of $o(n)$ using a Rectangular window and the biased formulation (no normalization by the number of samples used for the computation at each lag). This formulation was chosen in order not to favor values at large lags (which are usually estimated less precisely). We denote it by $X(l_p, t_i)$ where l_p represents the lags. $X(l_p, t_i)$ is normalized by its value at the zero-lag ($X(l_p = 0, t_i)$). It is therefore also independent of the local signal amplitude. $X(l_p > 0, t_i)$ reaches a maximum value of 1 only for highly periodic signals $o(n)$ around frame t_i . In the following we name it "ACF".

D. Product DFT and ACF (haDFTACF)

In [19] and [18], we have proposed a function which uses the fact that the octave uncertainty of the DFT and ACF occur in inverse domain (frequency domain for the DFT, lag domain or inverse frequency domain for the ACF). Because the combination of both DFT and ACF allows reducing octave errors, we have used it in [19] for pitch estimation and in [18] for tempo estimation. We test this function here in the case of rhythm description.

The combination of both functions is obtained by multiplying the values of the two functions after mapping one function to the domain of definition of the other: mapping the lags of the ACF to the frequencies of the DFT, or mapping the frequencies of the DFT to the lags of the ACF. For this, we compute at each frame t_i the DFT $A(k, t_i)$ and the ACF $X(p, t_i)$ on the same signal frame as explained above.

Product DFT / FM-ACF: Since, the value of the ACF at lag p represents the amount of periodicity at the lag $l_p = \frac{p}{sr}$ (where sr is the sampling rate) or at the frequency $f_p = \frac{1}{l_p} \forall l_p > 0$, l_p can be "mapped" in the frequency domain. We name Frequency-Mapped ACF (FM-ACF) the resulting function. In order to get the same linearly spaced frequencies f_k as for the DFT, we interpolate $X(l_p, t_i)$ and sample it at the lags $l_k = \frac{1}{f_k}$. We now have two measures (the DFT and the FM-ACF) of periodicity at the same frequencies f_k . We combined the two functions by computing their product at each frequency f_k : $P(f_k, t_i) = A(f_k, t_i) \cdot X(f_k, t_i)$.

Product TM-DFT / ACF: Inversely, since the value at the bin k of the DFT represents the amount of periodicity at

the frequency $f_k = k \frac{sr}{N}$ or at the lag $l_k = \frac{1}{f_k}$, f_k can be “mapped” in the lag domain. We name Temporally-Mapped DFT (TM-DFT) the resulting function. As before, we combine the functions by computing the product of the TM-DFT and the ACF at each lag l_p : $P(l_p, t_i) = A(l_p, t_i) \cdot X(l_p, t_i)$.

Product Hybrid Axis DFT / ACF (haDFTACF): Mapping the values of the ACF to the frequency domain f_k results in a loss of information. This is because when mapping the high lags l_p to the frequency domain, their initial spectral spacing ($\Delta f(p) = \frac{1}{l_{p-1}} - \frac{1}{l_p} = \frac{sr}{p-1} - \frac{sr}{p}$) is smaller than the constant resolution of the frequency axis ($\Delta f = f_k - f_{k-1} = \frac{sr}{N}$). This causes a sort of aliasing during the interpolation process. The same is true when mapping the DFT to the lag domain. The value of p for which the frequency resolution of the ACF is equal to the one of the DFT is given by $p_c = \frac{-1 + \sqrt{1 + 4N}}{2}$, where N is the number of points of the DFT. We therefore create an hybrid axis f_q made of the values of the lag axis mapped to the frequency domain $1/l_p$ for the lags $p < p_c$ and the values of the frequency axis f_k for $f_k > \frac{1}{l_{p_c}}$. Both the ACF and the DFT are mapped to this new axis and are interpolated at the new positions f_q . As before, we can combine the functions by computing the product at each frequency f_q : $P(f_q, t_i) = A(f_q, t_i) \cdot X(f_q, t_i)$.

It should be noted that the DFT/FM-ACF, TM-DFT/ACF and Hybrid Axis DFT/ACF only differ by the reference axis used for the mapping and the interpolation. Because the Hybrid axis DFT / ACF is the only one which does not suffer from aliasing, we only consider this one in this study. In the following we name it “haDFTACF”.

E. Other combinations of DFT and ACF coefficients (DFTandACF)

We also consider the direct concatenation of the DFT and ACF coefficients as a unique feature vectors. In this case, the machine-learning algorithms used in part IV will be able to use any combination of the ACF and DFT coefficients (such as summation) and not only the product of both as provided by the haDFTACF. In the following we name this concatenation “DFTandACF”.

F. Requirements

1) *Sensitivity to the sequence (order) of relative positions*: Among the representations mentioned in part I-B, neither the IOI histogram, nor the beat histogram are sensitive to the sequence of relative positions but only to the relative frequency of the duration². This was noticed by [9]. The authors take the example of a ChaCha pattern (which contains the following successive events JJJJJJ) and a Rumba pattern (JJJJJJ). They have different rhythm patterns but the same distribution of IOI and therefore the same IOI histogram (IOIH). This is illustrated in the top part of Fig. 3 ($\text{J}=1$ s. and $\text{J}=0.5$ s.). However, the amplitude of the DFT (hence the coefficients of the ACF and the ones of the haDFTACF) is sensitive to the sequence of relative positions.

²It should be noted however that the enhanced autocorrelation used for the calculation of the beat histogram is sensitive to the sequence of relative positions but not the beat histogram itself.

In the following we briefly explain how the temporal position of the events is encoded in the amplitude of the Fourier Transform through the phase. For this, we consider a simple model of a signal composed of a pulse train of period T and its Fourier Transform:

$$s_a(t) = \delta_T(t) = \sum_n \delta(t - nT) \leftrightarrow \frac{1}{T} \sum_n \delta(f - \frac{n}{T}) \quad (1)$$

The same signal shifted by Δ samples is expressed as

$$s_b(t) = \delta_T(t - \Delta) \leftrightarrow \frac{1}{T} \sum_n \delta(f - \frac{n}{T}) e^{-j\omega\Delta} \quad (2)$$

We now consider the sum of the two signals $s(t) = s_a(t) + s_b(t)$. If $\Delta = 0$ then $s(t)$ is a signal of period T , if $\Delta = \frac{T}{2}$ then $s(t)$ is a signal with period $\frac{T}{2}$. $s_a(t)$ and $s_b(t)$ correspond to two spectra with peaks of equal amplitude at the same frequencies (harmonics of $f_0 = \frac{1}{T}$) but with a de-phasing depending on Δ . If $\Delta = 0$ then $e^{-j\omega\Delta} = 1$ for all ω and the contribution of both signals are additive for all ω . If $\Delta \neq 0$ then the contribution of both signals are additive at the frequencies $f = \frac{k}{\Delta}$ and negative at the frequencies $f = \frac{(2k-1)}{2\Delta}$.

Meter. Following this, we represent a quadruple/simple meter pattern³ as $s_{42}(t) = A\delta_{4T}(t) + B\delta_T(t) + C\delta_T(t + \frac{T}{2})$ with $A > B > C > 0$ and a quadruple/compound meter pattern⁴ as $s_{43}(t) = A\delta_{4T}(t) + B\delta_T(t) + C\delta_T(t + \frac{T}{3}) + C\delta_T(t + \frac{2T}{3})$. From this formulation we can predict that the contributions of the terms of $s_{42}(t)$ are additive at the frequencies $f = k2f_0$ (negative at $f = (2k-1)f_0$) while for $s_{43}(t)$ they are additive at $f = k3f_0$ and $f = k\frac{3}{2}f_0$ (negative at $f = (2k-1)\frac{3}{2}f_0$ and $f = (2k-1)\frac{3}{4}f_0$). We illustrate the respective signals in Fig. 1, their corresponding IOIH, ACF, DFT and the Hybrid-Axis DFT-ACF. The predicted effects can be observed directly on the amplitude of the DFT peaks.

Swing. We model the introduction of a swing factor using the following model $s_{swing}(t) = A\delta_{4T}(t) + B\delta_T(t) + C\delta_T(t + \Delta)$ with $\Delta > \frac{T}{2}$. We illustrate this in Fig. 2 for two swing factors: $\Delta = \frac{2}{3}T$ and $\Delta = 0.7T$. It should be noted that when $\Delta = \frac{2}{3}T$ the signal corresponds to $s_{43}(t)$ without the term $C\delta_T(t + \frac{T}{3})$. This is reflected by a difference in the amplitude of the DFT peaks of $s_{43}(t)$ and $s_{swing}(t)$ (Fig. 1 right part and Fig. 2 left part). In the general case, the swing factor produces a modulation of the amplitude of the DFT peaks by $\cos(2\pi f\Delta)$. We also illustrate in Fig. 2 the influence of the swing factor on the other periodicity representations.

Pattern. The above mentioned ChaCha (JJJJJJ) and Rumba (JJJJJJ) rhythms can be modeled as: $s_{chacha}(t) = \delta_{4T}(t) + \delta_{4T}(t - T) + \delta_{4T}(t - 2T) + \delta_{4T}(t - 3T) + \delta_{4T}(t - 3T - \frac{T}{2})$ and $s_{rumba}(t) = \delta_{4T}(t) + \delta_{4T}(t - T) + \delta_{4T}(t - T - \frac{T}{2}) + \delta_{4T}(t - 2T - \frac{T}{2}) + \delta_{4T}(t - 3T - \frac{T}{2})$. We illustrate these patterns in Fig. 3. $s_{chacha}(t)$ contains the term $\delta_{4T}(t - 3T - \frac{T}{2})$ which is not in phase with the others. It creates the small troughs at 1 Hz, 3 Hz, 5 Hz (phase opposition). $s_{rumba}(t)$ contains three terms which are not in phase. They create stronger troughs in the DFT at 1 Hz, 3 Hz, 5 Hz (phase opposition) and small

³Each beat is sub-divided into two 8^{th} notes.

⁴Each beat is sub-divided into three 8^{th} notes.

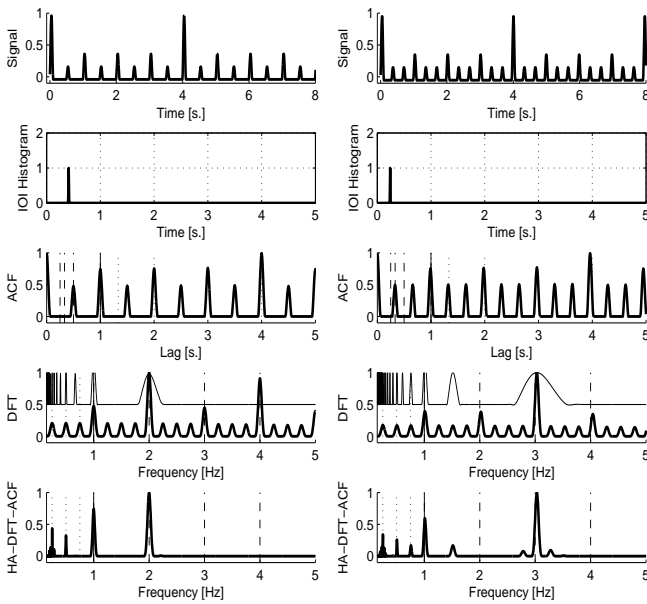


Fig. 1. Each panel represents from bottom to top: Signal, IOIH, ACF, DFT (super-imposed to it the FM-ACF) and haDFTACF. Vertical dashed lines on the ACF, DFT and haDFTACF represent $f_k = 1, 2, 3, 4$; Vertical dotted lines $f_k = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$. The signals have a tempo of 60bpm or 1Hz. [Left panel] quadruple/simple meter (each beat is divided into two 8^{th} notes), [Right panel] quadruple/compound meter (each beat is divided into three 8^{th} notes).

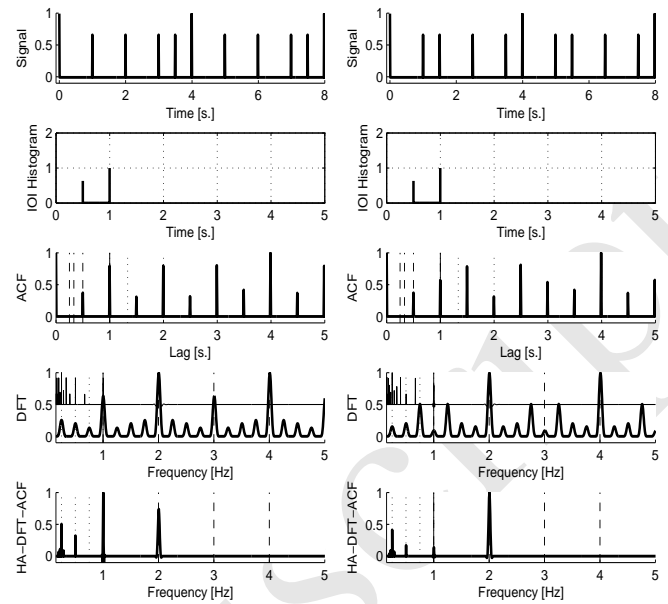


Fig. 3. Same as Fig. 1 but for [Left panel] ChaCha pattern, [Right panel] Rumba pattern.

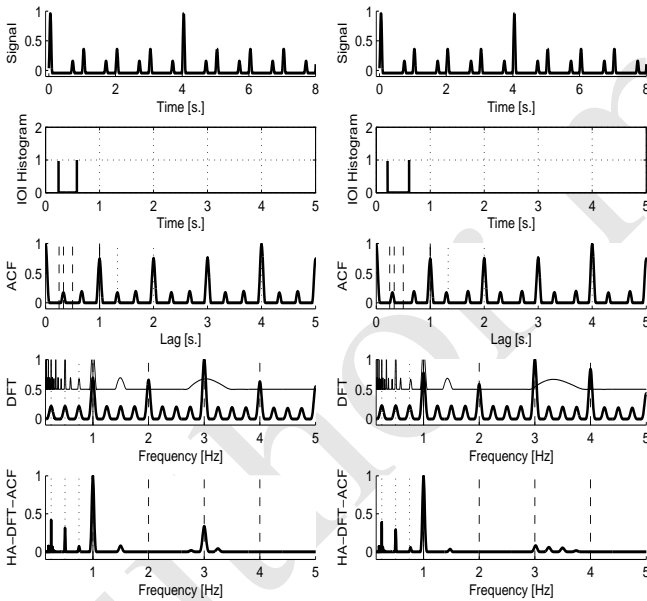


Fig. 2. Same as Fig. 1 but for [Left panel] Swing factor= 2/3%, [Right panel] Swing factor= 0.7%.

troughs at 0.5, 1.5, 2.5, 3.5 Hz (phase orthogonality). We also illustrate in Fig. 3 the influence of the pattern on the other periodicity representations. While the two IOIHs are identical, the ACFs and the haDFTACFs differ. This is especially true at lags 1.5, 2 and 2.5 for the ACF (the bar length is 4 sec), and at frequencies 0.5, 1 and 2 for the haDFTACF.

Discussion: The above formulations are valid whatever the distance between the observation time and the beginning of the sequence.

The DFT encodes the information on the rhythm pattern by the amplitude (and phase) at the frequencies which are at harmonic positions of the lowest common rhythm frequency, i.e. usually the bar frequency. The positions of these frequencies do not change when varying the meter, the rhythm pattern or when introducing swing in the rhythm (swing is present in the Quickstep and Jive classes of our test-set). While this fact is well known, it is important to insist on this since it allows us to use a fixed sampling-grid of the frequency axis. The DFT however necessitates a large number of coefficients to represent the rhythm pattern information.

The ACF also encodes the information on the rhythm pattern as the values at the lags which are at harmonic positions of the lowest common rhythm period, i.e usually the tatum period. However, in the presence of swing, the sampling of the lag axis is problematic because parts of the peaks of the ACF are not anymore in harmonic relationship with the tatum period.

The haDFTACF has the advantage (over the DFT and the ACF) to encode the information of the meter and rhythm pattern in only a few coefficients. However, as for the ACF, and because of the use of it, it faces the same problem in the presence of swing. The haDFTACF also face a “spectral leakage” problem: the values of the DFT at low frequencies are widened when interpolated to match the inverse-lag resolution⁵.

It should be noted that none of the techniques studied here is able to differentiate cyclic shifts of a sequence: such as differentiating the non-syncopated sequence ♩ · ♪♪♪ from the syncopated sequence ♪ · ♪♪♪ obtained simply by a cyclic shift of the first. This would require the estimation of a reference position (such as the downbeat position) that we did not consider here.

⁵It should be noted that we did not use peak-picking techniques in this study given the low reliability that these techniques have for non-peaky spectrum (as for the Slow-Waltz category of our test-set).

Remark: The three illustrations considered here (meter, swing, pattern) were made using a pulse train signal model. This has been chosen in order to make the formulations easier. It should be noted however that this model only roughly fits the onset-energy-function of a real music signal and does not fit it at all in many cases (music is not only based on periodic sequences). This is the reason why we introduce complementary features in part III. This model also relies on the assumption that the tempo remains constant within the observation window. While methods allowing to deal with frequency variations inside an observation window have been proposed (see for example [20]), we didn't find their use successful in the present context.

2) *Independence of the tempo:* In the following, we denote by $Y_f(f_k, t_i)$ the vector containing the values of the DFT, haDFTACF at the frequency f_k and time t_i , and $Y_l(l_k, t_i)$ the values of the ACF at the lag l_k and time t_i .

In order to make the feature vectors independent of the tempo, - we normalize the frequencies of $Y_f(f_k, t_i)$ by the local tempo frequency $f_{bpm}(t_i)$: $f'_k = \frac{f_k}{f_{bpm}(t_i)}$ or - we normalize the lags of $Y_l(l_k, t_i)$ by the local tempo period $l_{bpm}(t_i) = 1/f_{bpm}(t_i)$: $l'_k = \frac{l_k}{l_{bpm}(t_i)}$. Time-variable tempo is taken into account in our approach by performing the tempo-normalization at the frame-level. In the following we will use both annotated and estimated tempo $f_{bpm}(t_i)$.

From the tempo-normalized frame-values $Y_f(f'_k, t_i)$ (or $Y_l(l'_k, t_i)$), we then compute the mean value over time t_i . Each audio track is now represented by a single vector denoted by $Y_f(f'_k)$ (or $Y_l(l'_k)$).

3) *Compactness:* In order to reduce the size of the representation, we only retain from $Y_f(f'_k)$ (or $Y_l(l'_k)$) a reduced set of normalized frequencies selected to correspond to meaningful beat/tactus subdivision in tatum or grouping into bar. We consider the following sampling of the frequency axis

- For the DFT, we consider the normalized frequencies $f'_k \in V_f = \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1, 1.25, 1.33 \dots 8\}$. This is a 48-dimensional vector.
- For the ACF, we consider the normalized lags $l'_k \in V_l = \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1, 1.25, 1.33 \dots 8\}$. This is a 48-dimensional vector.
- For the haDFTACF, we consider the union of the normalized frequencies $f'_k \in V_f$ and inverse-lags $f'_k \in 1/V_l$. This is an 85-dimensional vector.
- When using the vector of concatenated DFT and ACF coefficients, we consider $f'_k \in V_f$ for sampling the DFT and $l'_k \in V_l$ for sampling the ACF. This is a 96-dimensional vector.

We denote by $Z(k')$ this reduced vector. In the case of V_f , the lower components (< 1) represent measure subdivision characteristics (energy at $\frac{1}{4}$, $\frac{2}{4}$ indicate binary meter; while energy at $\frac{1}{3}$ indicates ternary meter). The upper components (> 1) represent beat subdivision characteristics (1 represents energy at the 4th note, 2 the 8th note, 3 the 8th note triplet, 4 the 16th note). In the case of V_l , the lower components (< 1) represent beat subdivision characteristics and the upper components (> 1) measure subdivision characteristics.

In the following, we will also test automatic feature selection to further reduce the dimensionality of the feature vectors.

III. OTHER RHYTHM FEATURES

Because the sampled tempo-normalized periodicity representations $Z(k')$ do not represent all the characteristics of the rhythm (especially it does not allow to distinguish a highly periodic signal from a non-periodic signal), we propose here a set of complementary Rhythm Features⁶. We explain them here.

Mean bpm: is the average-over-track estimated tempo. We denote it by $RFbpmmean$.

Std bpm: is the standard-deviation-over-track of the estimated tempo. We denote it by $RFbpmstd$.

Meter is the meter estimated using the algorithm of [18]. Three choices are possible: 22 (binary grouping of the tactus into measure, binary subdivision into tatum), 23 (binary, ternary), 32 (ternary, binary). We denote it by $RFmeter$.

Percussivity The percussivity provides rough information on the amount of periodicity produced by percussive instruments. We define here a "percussive instrument" as a non-sustained instrument, i.e. with a fast decrease or a short "offset". While the use of Half-Wave-Rectification (HWR) applied to the time-derivative of the spectrogram bands highlights "onsets" of the signal, we use here a Negative-HWR (NHWR) to highlights "offsets". We then measure how much these "offsets" are responsible for the periodicity. The spectrogram $S(k, t)$ of the audio signal is computed using a hamming window of length 92.8ms and hop size of 5.8ms⁷. The time-derivative of the spectrogram is then computed and $NHWR(k, t)$ computed. $NHWR(k, t)$ is then grouped into four logarithmically spaced frequency bands b : $[\frac{sr}{32}, \frac{sr}{16}]$, $[\frac{sr}{16}, \frac{sr}{8}]$, $[\frac{sr}{8}, \frac{sr}{4}]$ and $[\frac{sr}{4}, \frac{sr}{2}]$ Hz⁸. We denote it by $NHWR_b(b, t)$. After subtracting its mean-value-over-time, the amplitude spectrum of $NHWR_b(b, t)$ is computed for each band. We denote it by $NHWR_f(b, f)$. We then compute two percussivity indexes:

- *RFperc1:* the energy of $NHWR_f(b, f)$ for f representing frequencies between 8.33 and 16.66 Hz (500 and 1000 bpm);
- *RFperc2:* RFperc1 normalized by the energy of $NHWR_f(b, f)$ between 0 and 8.33 Hz (0 and 500 bpm).

The contribution of all bands b are then summed up.

Periodicity The periodicity indicates how periodic the signal is. For this, using the same method as above, we compute $HWR_f(b, f)$ (instead of $NHWR_f(b, f)$). The global-slope-over-frequency (spectral envelope) of $HWR_f(b, f)$ is approximated using a second order-polynomial. Only the values of $HWR_f(b, f)$ above 1.5 this polynomial approximation

⁶These Rhythm Features are actually extracted using the same program as the one used for tempo estimation [18].

⁷These parameters are the same as the ones used in [18] for the computation of the onset-energy-function. As explained in [18], the choice of a large analysis window was made in order to increase the spectral resolution (separation between adjacent DFT peaks), it therefore allows an easier detection of onsets in the case of non-percussive instruments. However, the use of a large analysis window reduces the temporal resolution.

⁸This choice of frequency bands was motivated by the work of [21] for the computation of their Accent Filter Bank and by our own experiments. For a 11 KHz audio signal, these frequency bands correspond to [333, 678], [678, 1367], [1367, 2745] and [2745, 5500] Hz. It should be noted that we do not consider here the lowest frequency range which corresponds mainly to kick and bass-guitar sounds which we did not consider as percussive instruments.

are used in order to consider only salient frequencies. Peak-picking is then applied in order to retain only the components representing the energy of the periodic part of $HWR_f(b, f)$. We then compute two periodicity indexes:

- *RFperiod1*: it is the sum of the peak's energy of $HWR_f(b, f)$;
- *RFperiod2*: it is the ratio of the peak's energy contribution to the global energy of $HWR_f(b, f)$ for the frequencies between 0 and 16.66 Hz (0 and 1000bpm).

As before, the contribution of all bands b are then summed up.

Speed The "speed" gives information of the fastest musical event in the signal. This corresponds usually to the tatum periodicity. We compute the following two coefficients to represent it:

- *RFspeed1*: it is defined as the lag-centroid of the ACF values for lags between 0.166 and 13.33 Hz (10 and 800 bpm);
- *RFspeed2*: it is defined as the frequency corresponding to the first peak of the ACF with a value larger than 40% its maximum value.

It should be noted that these features are close to the ones proposed by [6].

IV. MUSIC GENRE CLASSIFICATION

In this part we compare the use of the various periodicity representations for a task of music genre classification.

A. Test-set

In order to be able to compare our results to previous approaches, we use the "Ballroom dancer" test-set [22]. This test-set is often used to measure the performances of rhythm characterization algorithms since it contains music for which the music genre and the rhythm genre are closely related. The "Ballroom dancer" test-set is composed of 698 tracks, each of 30 sec long, representing the following music genres: ChaCha (111 instances), Jive (60), QuickStep (82), Rumba (98), Samba (86), Tango (86), Viennese Waltz (65) and Slow Waltz (110).

B. Evaluation rules

In the following we compare the various periodicity representations, i.e. $Z(k')$ represents either the DFT (48 dimensions), ACF (48 dimensions), haDFTACF (85 dimensions) or DFTandACF (96 dimensions).

We test these periodicity representations computed using either annotated or estimated tempo $f_{bpm}(t_i)$. The estimated tempo is the one provided by the algorithm of [18]. It should be noted that we did not perform any optimization of our tempo estimation algorithm for this specific test-set. Our goal is actually to test the robustness of the periodicity representations in case of badly estimated tempo.

We also test three configurations of feature sets: - using only the periodicity representation, - using it combined with tempo information (annotated or estimated), - using it combined with tempo and the rhythm features mentioned in part III. The dimensionality of the rhythm feature vector is 9.

In order to allow the comparison of our results with the ones obtained in previous studies, we evaluate the performances using a ten-folds cross validation method. The results are presented as mean-accuracy over the ten-folds.

1) *Statistical hypothesis tests*: Considering that the values of the accuracy are only estimates (average over the ten-folds) of the real values, we perform a set of statistical hypothesis tests. Using the accuracy values obtained at each fold, we perform a set of pair wise Student T-tests with a 5% significance level. When comparing experiment A to B, we test the H_0 hypothesis that the mean-accuracy of A and B are equal, against the H_1 hypothesis that they are different. When the H_0 hypothesis is rejected (the mean values are statistically different) and when A is larger than B, we denote it by $A \gg B$. When the H_0 hypothesis is not rejected (there is not enough evidence to reject it), we denote it by $A \equiv B$. When comparing the results of several experiments A, B, and C we denote by $A \gg [B, C]$ the case $A \gg B$ and $A \gg C$. We denote by $[A, B] \gg [C]$ the case $A \gg C$ and $B \gg C$. We also use the acronym "SS" for "statistically significant".

2) *Classification algorithms*: For the classification, we use the following algorithms:

- J48: a C4.5. Decision tree algorithm [23],
- PART: a Partial Decision Tree algorithm [24],
- ClassViaReg: a Classification using regression methods [25],
- SVM: a Support Vector Machine with Polynomial kernel [26],
- AdaBoost: an Adaptive Boosting using a C4.5 Decision Tree [27],
- Random Forest: a Forest of random tree classifier [28].

For all these classifiers we have used the implementations provided by Weka [29] (version 3.6.2) with their default parameterization as provided by Weka; except for the AdaBoost, for which we have used a J48 decision tree as weak classifier.

C. Features visualization

In Fig. 4 we represent the periodicity representation $Y_f(f'_k)$ in the case of the haDFTACF and when using annotated tempo for its estimation for all the songs belonging to each music genre of the "Ballroom dancer" test-set. The left part represents the values in linear-frequency-scale for all the songs of a given class, the right part represents the average-over-track $Y_f(f'_k)$ in logarithmic-frequency-scale for better visualization; we super-impose over it the positions $f'_k = 1/4, 1/2, 1, 2, 3, 4$ (continuous vertical lines), and $f'_k = 1/3, 2/3$ (dashed vertical lines). Some characteristics of music genre appear immediately on this representation: Viennese-Waltz and Slow Waltz are the only genres having a component at $f'_k = 1/3$ (3/4 meter), but Viennese-Waltz has no (a weak) component at $f'_k = 2$ (8th note) while Slow-Waltz has, Samba is the only genre having a component at $f'_k = 4$ (16th note), Jive and QuickStep have no component at $f'_k = 2$.

In Fig. 5, we represent the annotated and estimated tempo distribution for the eight musical genres of the test-set. Jive and Quickstep are systematically estimated one-octave below the ground-truth tempo, Viennese-Waltz is most of the time

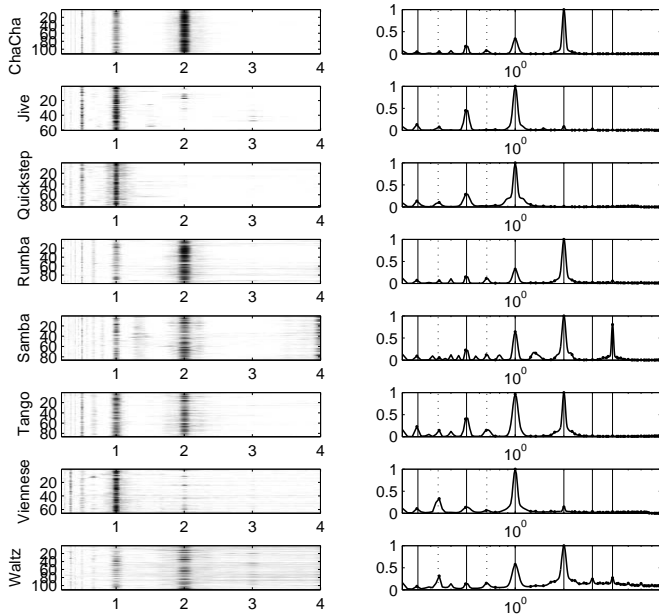


Fig. 4. Periodicity representation $Y_f(f'_k)$ (using haDFTACF) using annotated tempo for the various music genres of the “Ballroom dancer” test-set. In each plot, the x-axis represents f'_k , the y-axis the track number.

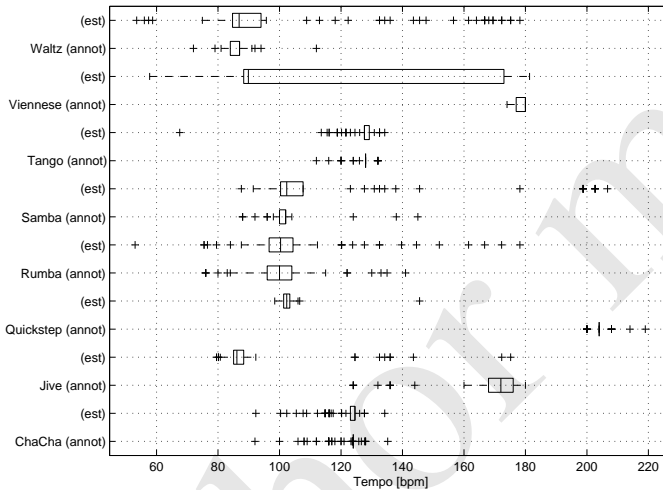


Fig. 5. Distribution (in the form of box and whisker plot) of annotated and estimated tempi for the eight musical genres of the “Ballroom dancer” test-set. For each genre, the lower box corresponds to annotated tempi (annot), the upper one to estimated tempi (est).

estimated one octave below (considering that Viennese-Waltz is in 3/4, the estimated tempo correspond to the ♩).

D. Results and discussion

The recognition rates obtained using the four feature sets (with and without tempo information, with and without rhythm features, using annotated or estimated tempo) are indicated in Table I. Below each accuracy, we indicate the best classifiers used. In almost all cases, the best results were obtained using Support Vector Machine or AdaBoost.

Using only tempo allows achieving 80.8% (annotated tempo, **Ta**) and 65.2% (estimated tempo, **Te**) accuracy. Using

TABLE I

ACCURACY USING VARIOUS PERIODICITY REPRESENTATIONS (DFT, ACF, haDFTACF, DFTANDACF) USED ALONE, WITH TEMPO INFORMATION, WITH TEMPO AND RHYTHM FEATURES. THE LOWER PART OF THE TABLE REPRESENTS THE RESULTS OBTAINED WHEN USING AUTOMATIC FEATURE SELECTION. RESULTS ARE INDICATED IN TERMS OF ACCURACY. BELOW EACH RESULT WE INDICATE THE BEST CLASSIFIER USED. IN THE CASE OF AUTOMATIC FEATURE SELECTION, WE ALSO INDICATE THE BEST NUMBER OF FEATURES: 38-D SVM MEANS “38-DIMENSIONAL FEATURE VECTOR” WITH A SVM CLASSIFIER.

Accuracy	Annotated			Estimated		
	Alone	+ Bpm (Ta)	+ Bpm (Ta) + Features (RF)	Alone	+ Bpm (Te)	+ Bpm (Te) + Features (RF)
BPM (Ta/Te)	80,8%			65,2%		
best classifier	J48			Part		
Features (RF)	80,2%			80,2%		
best classifier	AdaBoost			AdaBoost		
BPM + Features	90,3%			80,4%		
best classifier	AdaBoost			AdaBoost		
DFT	93,4%	94,8%	95,6%	80,1%	84,1%	86,5%
best classifier	SVM	SVM	SVM	SVM	AdaBoost	AdaBoost
ACF	80,1%	92,7%	93,1%	68,3%	80,7%	84,2%
best classifier	SVM	AdaBoost	AdaBoost	AdaBoost	ClassViaPet	AdaBoost
HA-DFT-ACF	83,8%	91,8%	92,3%	65,6%	81,8%	83,7%
best classifier	SVM	AdaBoost	AdaBoost	ClassViaPet	AdaBoost	AdaBoost
DFT AND ACF	93,7%	94,6%	95,3%	83,8%	84,2%	86,2%
best classifier	SVM	SVM	SVM	SVM	SVM	AdaBoost
DFT	93,98	95,41	96,13	80,94	85,53	87,1
best dim/classifier	38-d SVM	26-d SVM	40-d SVM	40-d SVM	18-d AdA	18-d AdA
DFT AND ACF	94,26	94,84	95,7	83,81	85,81	87,96
best dim/classifier	28-d SVM	30-d SVM	36-d SVM	initial SVM	34-d AdA	28-d AdA

only Rhythm Features **RF** (which do not use at all annotated tempo) allows achieving 80.2% accuracy. Combining tempo and Rhythm Features information allows achieving 90.3% (annotated tempo) and 80.4% (estimated tempo) accuracy.

The hypothesis tests lead to $[Ta+RF] \gg [Ta, RF]$ and $[RF, (Te+RF)] \gg [Te]$. It means that in the case of annotated tempo, the simultaneous use of Ta and RF leads to a SS improvement of the accuracy. In the case of estimated tempo, the use of RF (alone or combined with Te) leads to a SS improvement of the accuracy.

1) *Using annotated tempo:* When considering annotated tempo, the best periodicity representations are the DFT (93.4%) and the DFTandACF (93.7%). The hypothesis tests lead to $[DFT, DFTandACF] \gg [ACF, haDFTACF]$ and $[DFT] \equiv [DFTandACF]$. The DFT and DFTandACF are therefore SS better than the ACF and the haDFTACF for the given task of classification but statistically equal between each other.

The accuracy reaches 95.6% (DFT) and 95.3% (DFTandACF) when combined with tempo and Rhythm Features information. Also, in this case, $[(DFT+Ta+RF), (DFTandACF+Ta+RF)] \gg [(ACF+Ta+RF), (haDFTACF+Ta+RF)]$. It means, that even when combined with tempo and Rhythm features information, the DFT and DFTandACF are SS better than the ACF and the haDFTACF for the given task of classification. There is however no SS differences between the $(DFT+Ta+RF)$ and the $(DFTandACF+Ta+RF)$. There is also no SS differences between the results obtained with the

TABLE II
 CONFUSION MATRIX USING ONLY DFTANDACF FEATURES COMPUTED FROM ANNOTATED TEMPO

Classified as ->	ChaCha	Jive	QuickStep	Rumba	Samba	Tango	Viennese Waltz	Slow Waltz	
ChaCha	100			6		5			111
Jive	1	56	1			2			60
QuickStep			81				1		82
Rumba	9			82		2		5	98
Samba	1			3	80	1		1	86
Tango	1	1		2		82			86
Viennese Waltz							65		85
Slow Waltz								108	110
Precision	89,3%	98,2%	98,8%	86,3%	100,0%	89,1%	98,5%	94,7%	

Recall
 90,1%
 93,3%
 98,8%
 83,7%
 93,0%
 95,3%
 100,0%
 98,2%

Accuracy
 93,70%

DFT or DFTandACF alone (93.4% and 93.7%) or when combined with tempo and Rhythm Features information (95.6% and 95.3%): [(DFT+Ta+RF), (DFTandACF+Ta+RF)] ≡ [DFT, DFTandACF]. It means that the increase of accuracy provided by the tempo and Rhythm Features information is not SS.

2) *Using estimated tempo:* When considering estimated tempo, the best periodicity representation is the DFTandACF (83.8%). Also in this case [DFT, DFTandACF] ≫ [ACF, haDFTACF] but there is no SS differences between the DFT and the DFTandACF: [DFT] ≡ [DFTandACF].

The accuracy reaches 86.2% (DFTandACF) and 86.5% (DFT) when combined with tempo and Rhythm Features information. However, in this case, no SS differences were found between the four representations: DFT ≡ ACF ≡ haDFTACF ≡ DFTandACF. For the DFT, ACF and haDFTACF, the introduction of the estimated tempo Te and the Rhythm Features RF leads to a SS increase of accuracy: [DFT+Te+RF] ≫ [DFT] also [ACF+Te+RF] ≫ [ACF] and [haDFTACF+Te+RF] ≫ [haDFTACF]. This is not the case for the DFTandACF: [DFTandACF+Te+RF] ≡ [DFTandACF].

All the results obtained using annotated tempo are also SS higher than their equivalent using estimated tempo.

As expected from our previous discussions (on the influence of the swing factor), the ACF and the haDFTACF performances are below the ones of the DFT and the DFTandACF. However, the bad performances of the ACF (in the presence of swing) seem not to be propagated to the DFTandACF.

3) *Confusion matrix:* In Table II we indicate the confusion matrices corresponding to the DFTandACF case when used alone (without tempo and Rhythm Features information) for the case of annotated tempo (accuracy of 93.7%). If we denote by \hat{c} the estimated class of c , the largest confusions are - \hat{ChaCha} =Rumba, Tango, - \hat{Rumba} =ChaCha, Slow-Waltz, - \hat{Samba} =Rumba, - \hat{Tango} =Rumba. These larger confusions can be explained either by their close tempi (see Fig. 5) or their close periodicity representation $Z(k')$ (see Fig. 4 for the case $Z(k') = haDFTACF$).

In the case of estimated tempo, the same confusions are observed but also - \hat{Jive} =QuickStep, Tango, - $\hat{Quickstep}$ =Jive, Rumba. There is also some confusion between Viennese-Waltz and Slow-Waltz: 12/65 tracks of Viennese-Waltz are recognized as Slow-Waltz, 9/110 tracks of Slow-Waltz are recognized as Viennese-Waltz.

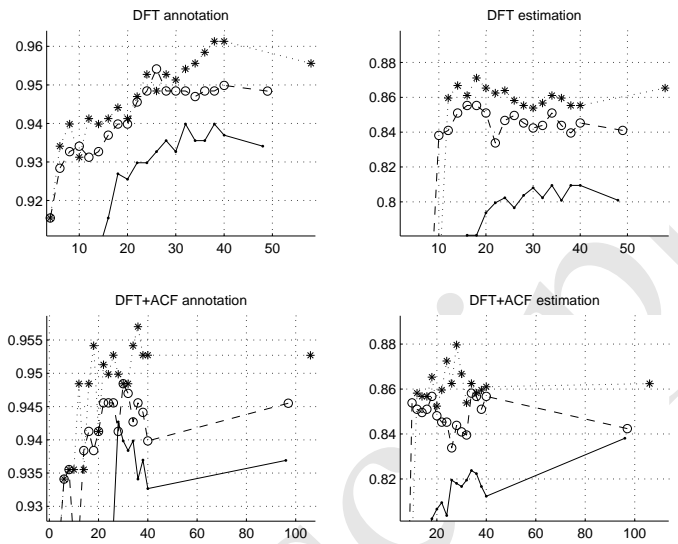


Fig. 6. Evolution of accuracy (y-axis) when reducing the number of features (x-axis) using feature selection for the DFT and DFTandACF feature set using annotated (left) and estimated tempo (right): (—) using only the periodicity representation, (o—) combined with the tempo information (*··) combined with the tempo and rhythm features information.

E. Best features

In order to better understand the discriminative power of each element k' of $Z(k')$, we have applied an **automatic feature-selection (AFS)**. The algorithm we have used is our Inertia Ratio Maximization with Feature Space Projection (IRMFSP) algorithm⁹. In [30], we positively compared the IRMFSP algorithm to the most-used CFS [31] algorithm.

For the DFT and DFTandACF case, we redo the same experiment as before (testing for periodicity representation alone, with tempo and with rhythm features information, using either annotated or estimated tempo and testing with the six different classifiers) reducing the number of features to the 40 best ones, 38, 36 ... down to 4). The selection concerns both the dimension of the pattern $Z(k')$, the use or not of tempo information and of each rhythm feature. The evolution of the accuracy over the number of features is indicated in Fig. 6. The accuracy obtained using the whole set of features is represented by the rightmost points of each plot. It should be noted that the accuracy of each point has been potentially obtained using a different classifier. In the lower part of Table I, we indicate the best number of features and the corresponding classifier for each case.

In these plots, we observe the well-known fact that reducing the number of features to only the relevant ones can improve the recognition results. For example, when using annotated tempo, the best result was 95.6% with DFT + Tempo + Rhythm-Features using SVM. It now reaches **96.13%** using DFT and only 40 features using SVM. When using estimated tempo, the best result was 86.5% with DFT + Tempo +

⁹The IRMFSP algorithm is an iterative algorithm which selects one feature at a time. The selection is based on the feature with the largest Fisher ratio. It then performs a Gram-Schmidt orthogonalization of the whole feature space on the selected feature. This guarantees that the remaining features are no longer correlated with the selected features. The selection process is then repeated until the Fisher ratio passes under a specific threshold.

Rhythm-Features using AdaBoost. It now reaches **88%** using DFTandACF and only 28 features using AdaBoost. However, these differences are not SS.

We indicate this set of 28 features (in order of selection): RFperiod1, $ACF(l'_k=4.75)$, $DFT(f'_k=6)$, $ACF(l'_k=0.25)$, $DFT(f'_k=1.75)$, $DFT(f'_k=8)$, RFmeter, $ACF(l'_k=7.5)$, $DFT(f'_k=5.666)$, $ACF(l'_k=8)$, $ACF(l'_k=6)$, RFbpmmean, $DFT(f'_k=7.5)$, $DFT(f'_k=0.75)$, $DFT(f'_k=0.666)$, $ACF(l'_k=1)$, $DFT(f'_k=0.5)$, RFperc1, $ACF(l'_k=5.5)$, $ACF(l'_k=1.5)$, $DFT(f'_k=2)$, $DFT(f'_k=4)$, RFspeed1, RFperc2, $ACF(l'_k=4.333)$, $ACF(l'_k=6.333)$, $ACF(l'_k=7)$, $DFT(f'_k=1.5)$.

F. Comparison with results published in previous studies

96.13% is close to the best results obtained using annotated tempo by [9] (96%). 88% is above the results obtained without using annotated tempo by [12] (86.9% with the Melin Transform) or by [9] (85.7%).

V. CONCLUSION AND FUTURE WORKS

In this paper, we have studied the use of four periodicity representations to describe the rhythm characteristics of an audio item. For a task of music genre classification, we have shown that the use of simple representation such as the DFT or the concatenated DFTandACF allows achieving high recognition rates at least comparable to previously published results (obtained with more complex methods). When considering annotated tempo, the best result (96.13%) is obtained using a 40-dimensional feature vector composed of sampled values of the DFT, the tempo and a small set of rhythm features. In this case, using the sample values of the DFT alone achieves 94% accuracy. When considering estimated tempo, the best result (87.96%) is obtained using a 28-dimensional feature vector composed of sampled values of the concatenated DFTandACF, the tempo and a small set of rhythm features. In this case, using the sample values of the ACF-and-DFT alone achieves 84% accuracy. Considering the simplicity of our approach, these results are promising.

While we have shown in our previous studies that the product DFT and ACF function, thanks to its good periodicity discrimination properties, allows improving pitch or tempo estimation; in the present case this discrimination is not beneficial. As explained, part of the problem comes from the peak localization of the ACF in the presence of swing. Another problem comes from the widening of spectral lobes occurring during the interpolation process. A potential solution could be to add a peak-picking process or to use the narrowed-ACF proposed by [32].

The aim of this paper was to provide a quick and efficient (robust) way to perform classification of rhythm into classes. For these reasons, we did not study the use of more elaborate periodicity measures based on the assumption of a possible sinusoidal representation (such as peak-picking, frequency-reassignment or least-square sinusoidal model parameter estimation). For computational efficiency reasons, we also did not consider the use of more elaborate way of comparing the periodicity representations such as Dynamic Programming or Scale/ Melin transform such as proposed by [11][12].

While we have shown that simple representations such as the amplitude of the DFT allow representing efficiently the rhythm content of an audio item, part of the temporal organization of the rhythm is however not represented in it. Since our representation does not use a reference position (such as the downbeat positions), it is not able to distinguish between cyclic shifts of a sequence. Part of the missing information is actually contained in the phase spectrum of the DFT. Another part is contained in the frequency localization of the events forming the rhythm (low / high frequency). In the present study only a single onset-energy-function was used to represent the whole audio signal without differentiating the spectral role of the various events. Finally, not all information can be derived from an onset-energy-function and higher-level information such as harmonic changes over time should be considered. All these points will be the subject of our future works. We would also like to extend our evaluation to a larger set of music genre and test our approach for the recognition of additive meters.

ACKNOWLEDGMENT

We would like to thank the three anonymous reviewers for their fruitful comments which help in the improvement of this paper. Part of this work was conducted in the context of the European IST project Semantic HIFI and Quaero Oseo French Project.

REFERENCES

- [1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [2] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proc. of ICME (IEEE Int. Conf. on Multimedia and Expo)*. Pal Xerox FXPAL-PR-01-022, 2001, pp. 1088–1091.
- [3] J. Foote, M. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," in *Proc. of ISMIR*, Paris, France, 2002, pp. 265–266.
- [4] I. Antonopoulos, A. Pikrakis, S. Theodoridis, O. Cornelis, D. Moelants, and M. Leman, "Music retrieval by rhythmic similarity applied on greek and african traditional music," in *Proc. of ISMIR*, Vienna, Austria, 2007, pp. 297–300.
- [5] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. of ISMIR*, Paris, France, 2002, pp. 150–156.
- [6] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proc. of AES 25th Int. Conf. on Metadata for Audio*, London, UK, 2004, pp. 196–204.
- [7] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," in *Proc. of ISMIR*, Baltimore, Maryland, USA, 2003, pp. 201–208.
- [8] F. Gouyon and S. Dixon, "Dance music classification: a tempo-based approach," in *Proc. of ISMIR*, Barcelona, Spain, 2004, pp. 501–504.
- [9] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *Proc. of ISMIR*, Barcelona, Spain, 2004, pp. 509–516.
- [10] M. Wright, W. Schloss, and G. Tzanetakis, "Analyzing afro-cuban rhythms using rotation-aware clave template matching with dynamic programming," in *Proc. of ISMIR*, Philadelphia, PA, USA, 2008, pp. 647–652.
- [11] A. Holzapfel and Y. Stylianou, "Rhythmic similarity of music based on dynamic periodicity warping," in *Proc. of IEEE ICASSP*, Las Vegas, USA, 2008.
- [12] —, "A scale transform based method for rhythmic similarity of music," in *Proc. of IEEE ICASSP*, Taipei, Taiwan, 2009.
- [13] J. Jensen, M. Christensen, and S. Jensen, "A tempo-insensitive representation of rhythmic patterns," in *Eusipco*, Glasgow, Scotland, 2009.
- [14] M. Gruhne, C. Dittmar, and D. Gaertner, "Improving rhythmic similarity computation by beat histogram transformations," in *Proc. of ISMIR*, Kobe, Japan, 2009.

- [15] C. Uhle and C. Dittmar, "Drum pattern based genre classification of popular music," in *Proc. of AES 25th Int. Conf. on Metadata for Audio*, London, UK, 2004.
- [16] E. Tsunoo, G. Tzanetakis, N. Ono, and S. Sagayama, "Audio genre classification using percussive pattern clustering combined with timbral features," in *Proc. of ICME (International Conference on Multimedia and Expo)*, 2009, pp. 382–385.
- [17] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. of ISMIR*, Philadelphia, USA, 2008.
- [18] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 158–158, 2007.
- [19] —, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. of IEEE ICASSP*, vol. V, Toulouse, France, 2006, pp. 53–56.
- [20] G. Peeters and X. Rodet, "Non-stationary analysis/synthesis using spectrum peak shape distortion, phase and reassignment," in *Proc. of ICSPAT - DSP World*, D. World, Ed., Orlando, Florida, USA, 1999.
- [21] J. Seppanen, A. Eronen, and J. Hiipakka, "Joint beat and tatum tracking from music signals," in *Proc. of ISMIR*, Victoria, Canada, 2006.
- [22] Ballroom-Dancers.com.
- [23] J. R. Quinlan, *C4.5.: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann, 1993.
- [24] E. Frank and I. Witten, "Generating accurate rule sets without global optimization," in *Proc. of ICML (Int. Conf. on Machine Learning)*, 1998, pp. 144–151.
- [25] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. Witten, "using model trees for classification," *Machine Learning*, vol. 32, no. 1, pp. 63–76, 1998.
- [26] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–127, 1998.
- [27] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119–139, 1997.
- [28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [29] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Fransisco, CA: Morgan Kaufmann, 1999.
- [30] G. Peeters and X. Rodet, "Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instrument database," in *Proc. of DAFX*, London, UK, 2003, pp. 318–323, peeters03c.
- [31] M. Hall, "Feature selection for discrete and numeric class machine learning," Tech. Rep., 1999.
- [32] J. Brown and M. Puckette, "Calculation of a "narrowed" autocorrelation function," *J. Acoust. Soc. Am.*, vol. 85, no. 4, pp. 1595–1601, 1989.



Geoffroy Peeters Geoffroy Peeters received 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. He has also coordinated indexing research activities for the Cuidad, Cuidado, and Semantic HIFI European projects and is currently leading the music indexing activities in the Quaero Oseo project. He is one of the co-authors of the ISO MPEG-7 audio standard.

description, automatic music structure discovery, and audio summary. He owns several patents in these fields. He has also coordinated indexing research activities for the Cuidad, Cuidado, and Semantic HIFI European projects and is currently leading the music indexing activities in the Quaero Oseo project. He is one of the co-authors of the ISO MPEG-7 audio standard.