RHYTHM CLASSIFICATION USING SPECTRAL RHYTHM PATTERNS

Geoffroy Peeters
IRCAM - Sound Analysis/Synthesis Team
1, pl. Igor Stravinsky
75004 Paris - France
peeters@ircam.fr

ABSTRACT

In this paper, we study the use of spectral patterns to represent the characteristics of the rhythm of an audio signal. A function representing the position of onsets 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. Three feature sets are studied for this vector. They are derived from the amplitude of the Discrete Fourier Transform, the Auto-Correlation Function and the product of the DFT and of a Frequency-Mapped ACF. The vectors are then sampled at some specific frequencies, which represents various ratios of the local tempo. The ability of the three feature sets to represent the rhythm characteristics of an audio item is evaluated through a classification task. We show that using such simple spectral representations allows obtaining results comparable to the state of the art.

Keywords: rhythm representation, classification

1 INTRODUCTION

Automatic music description from signal analysis has become one of the major 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.

For this representation, 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 acoustical characteristics of the events or both) and the way they are represented (sequence of events, histogram, profiles, evolution, ...). [2] proposes the use of a *beat spectrum* (obtained by summing the signal similarity matrix along diagonals at specific lags) to visualize the temporal structure of a song (beat, measure and small structure). [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. [3] proposes to model the rhythm characteristics as a sequence of audio features (loudness, spectral centroid, ...) along time. A Dynamic Time Warping algorithm is then used to align time and allows the comparison of two sequences of different lengths. Gouyon's work is also based on audio features. [4] tests a set of 73 features to characterize the rhythm. These include features derived from the tempo, from a *periodicity histogram* and from the *Inter-Onset-Interval Histogram* (IOIH). These features are used for the classification of 8 music genres from the "ballroom dancer" database. The authors report 90.1% correct recognition using the correct tempo (78.9% using the estimated tempo). Another study made by Gouyon [5] considers tempo estimation errors as part of the estimation process. They use 28 pair-wise classifiers and obtain 67.6% correct recognition. A recent study by Dixon [6] proposes to add to Gouyon set of features, a representation of the *temporal rhythmic patterns* derived from the energy evolution of the signal inside each bar. This pattern represents the temporal position of the events. Various other features are also used (meter, syncopation, swing factor, ...). The performances are also tested on the "ballroom dancer" database. The authors report 50% correct recognition using only the pattern, and up to 96% using this pattern and all features with an AdaBoost classifier.

In this paper, we study the use of three simple spectral patterns to characterize the rhythm. The paper is organized as follows. In part 2.1, we give a quick overview of our global tempo estimation system. In part 2.2., we propose the three spectral rhythm patterns. In part 3, we compare the use of these representations in a task of music genre classification and compare our results with the state of the art.

2 PROPOSED METHOD

In [7] we have proposed a system for the estimation of the tempo and beat positions of a piece of music. This system is the basis for our spectral rhythm representation.

2.1 Tempo estimation

Onset estimation: An onset-energy function is first extracted from the audio signal. In order to allow a robust detection of onsets even in case of music with non-percussive instruments we propose the use of a reassigned spectral energy flux obtained from the reassigned spectrum [8].

Periodicities estimation: The onset-energy function is then used to estimate the dominant periodicities at a given time. This could be done using either Discrete Fourier Transform (DFT) or AutoCorrelation Function (ACF), but we propose the use of a combination of DFT and Frequency-Mapped ACF. *Why using both the DFT and the ACF?* The DFT of a periodic signal is a set of harmonically related frequencies. Depending on their relative amplitude it can be difficult to decide which one of the harmonics corresponds to the tempo frequency. This ambiguity can lead to octave errors which are especially detrimental in the case of triple or compound meter (in these cases octave errors can lead to musically insignificant frequencies). The same occurs for the ACF but in the time domain. Because 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), we use this property to construct a product function that reduces these ambiguities. *Calculation:* • At each frame t_i, the DFT $F(\omega_k, t_i)$ and the ACF $A(l, t_i)$ are computed on the same signal frame[1]. • The value at lag l of the ACF represents the amount of periodicity at the lag l/sr (where sr is the sampling rate) or at the frequency $\omega_l = sr/l \ \forall l > 0$. Each lag l is therefore "mapped" in the frequency domain. • In order to get the same linearly spaced frequencies ω_k as for the DFT, we interpolate $A(l, t_i)$ and sample it at the lags $l = sr/\omega_k$. • We now have two measures (the DFT and the FM-ACF) of periodicity at the same frequencies ω_k. We finally combined the functions by computing the product of the DFT and the FM-ACF at each frequency ω_k: $Y(\omega_k, t_i) = F(\omega_k, t_i) \cdot A(\omega_k, t_i)$.

In Figure 1 we illustrate the interesting properties for rhythm characterization of this product function for two signals at 120 bpm: - a simple meter in 4/4 (each beat is divided into 8^{th} note), - a compound meter in 4/4 (each beat is divided into 8^{th} note triplet). We represent the mean value over time of the DFT, the ACF and the product function[2]. The product function allows to better emphasize the 4^{th} note / 8^{th} note frequencies for the simple meter, the 4^{th} note / 8^{th} note triplet frequencies for the compound meter therefore reducing octave ambiguities. Spurious peaks however exist due to the spectral leakage and the frequency resampling process of the ACF.

Tempo estimation: In [7], we have derived the most likely tempo path over time $\omega_{bpm}(t_i)$ from $Y(\omega_k, t_i)$ using a Viterbi decoding algorithm. However in this paper, we are only interested in the discriminative power of $Y(\omega_k, t_i)$ for rhythm characterization, apart from the precision of the tempo estimation. Because of that, in the rest of the paper we will use a ground-truth tempo.

[1]We use a window length of 6 s. and a hop size of 0.5 s. For a simple meter in 4/4 at 120bpm, this allows the observation of 3 measures of 4 beats with a good spectral resolutions.

[2]The window length was set to 3 s. and the hop size to 0.5 s.

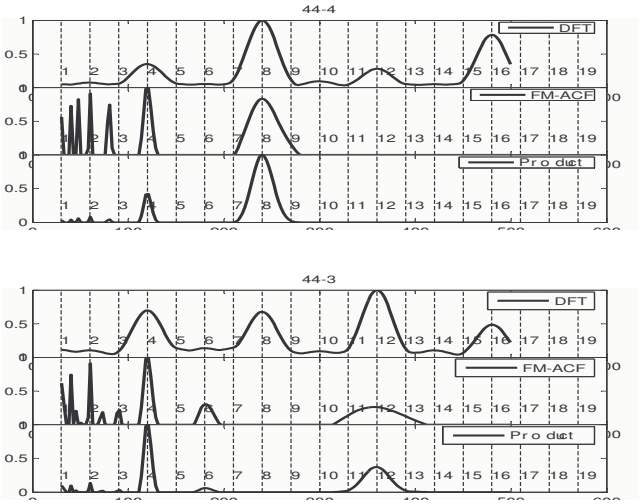

Figure 1: Comparison of DFT, FM-ACF and product function for [top three] simple meter in 4/4 [bottom three] compound meter in 4/4; x-axis: frequency, y-axis: amplitude, dotted lines denote the frequency of symbolic durations (4: $4th$ note, 8: 8^{th} note, 12: 8^{th} note triplet, ...)

2.2 Spectral rhythm patterns

Rhythm can be roughly defined by the tempo, the position and duration of the events and their acoustical characteristics. Instead of the position of the events, we usually prefer to work on the sequence of event's duration (or the successive Inter-Onset-Intervals). In this paper, we are only interested in the representation of the sequence of duration not on the acoustical characteristics of the events. We seek a representation of the rhythm that is • sensitive to the sequence (order) of duration (but robust to small changes) • independent of the tempo (i.e. the speed of reading of the sequence) • compact.

Sensitiveness to the sequence (order) of duration: Among the representation mentioned in part 1, neither the IOI histogram, nor the beat histogram are sensitive to the sequence of duration but only to the relative frequency of the duration. This was noticed by [6]. The authors take the example of a ChaChaCha pattern (which contains the following successive events ♩♩♩♪♪) and a Rumba pattern (♩♪♩♩♪). 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 Figure 2 (♩=0.5 s. and ♪=0.25 s.). However, the amplitude of the DFT is sensitive to the sequence of duration through the phase relations[3]. Since the ACF is the inverse of the power spectrum, it is also sensitive to the sequence of duration, and therefore also the product DFT/FM-ACF. This is illustrated in the remaining part of Figure 2. While the two IOIHs are identical, the ACFs, DFTs and product DFT/FM-ACF of both patterns differ. Considering

[3]A simple mathematical demonstration of this can be made by representing the above-mentioned signals as the summation of pulses trains of period $4T$ (T being the tempo period) with various time-shifts. Considering that these time-shifts (Δ) introduce phase modifications in the complex spectrum ($e^{-j\omega\Delta}$), and therefore influence the addition of the components in the complex domain (components in phase, in phase opposition, ...).

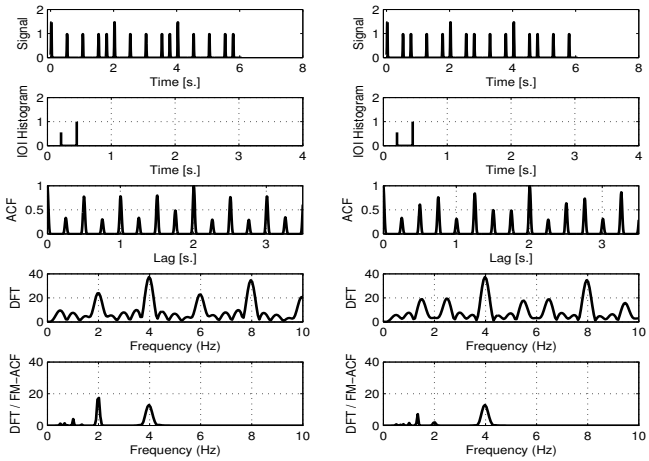

Figure 2: [Left column] ChaChaCha pattern and [right column] Rumba pattern represented by [from top to bottom] Temporal pattern, IOI Histogram, ACF, DFT and product DFT/FM-ACF

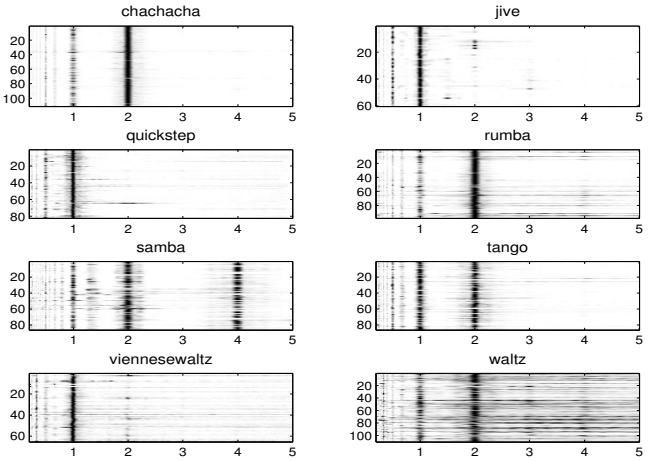

Figure 3: Spectral rhythm patterns $\bar{Y}_n(\omega'_k)$ (using product DFT/FM-ACF) for the various music genres of the "ballroom dancer" database (x-axis: normalized frequencies, y-axis: item's number on each category).

that, we propose these three functions as candidates feature vectors for rhythm characterization.

Independence of the tempo: We note $Y(\omega_k, t_i)$ either the DFT, the ACF or the product DFT/FM-ACF vector at the frequency ω_k and time t_i. • For each track, we extract the series of vectors $Y(\omega_k, t_i)$. • In order to make the feature sets independent of the tempo, we normalize the frequencies of $Y(\omega_k, t_i)$ by the local tempo frequency $\omega_{bpm}(t_i)$: $\omega'_k = \frac{\omega_k}{\omega_{bpm}(t_i)}$. • We then compute the mean of $Y(\omega'_k, t_i)$ over time t_i. • Finally, the vector is normalized to unit sum. Each track is now represented by a single vector $\bar{Y}_n(\omega'_k)$.

Compactness: We ony retain from $\bar{Y}_n(\omega'_k)$ a reduced set of normalized frequencies selected to correspond to musically meaningfull frequency: $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4. The lower components (< 1) represent measure subdivision characteristics. The upper components (> 1) represent beat subdivision characteristics. The reduced vector is noted $Z(k)$ and is called a *spectral rhythm pattern*. It is a pattern which represent the amount of energy at musically meaningfull frequencies.

3 MUSIC GENRE CLASSIFICATION

In this part we compare the use of the DFT, the ACF and the product DFT/FM-ACF functions for the task of music genre classification.

Data: As in [5] and [6], we use the "ballroom dancer" database [9] because this database contains music genres for which there is a close link between the music genre and the rhythm genre. The "ballroom dancer" database is composed of 698 tracks, each of 30 sec long, representing the following music genre: ChaChaCha (111 instances), Jive (60), QuickStep (82), Rumba (98), Samba (86), Tango (86), Viennese Waltz (65), Waltz (110).

Features: In the following we compare the three feature sets derived from the DFT, the ACF and the product DFT/FM-ACF functions. In each case, we consider the use of each feature set alone and the use of it combined

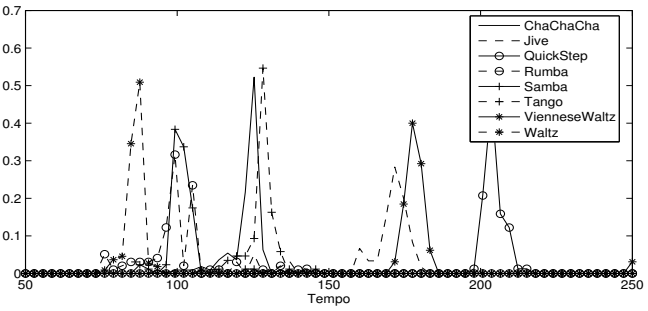

Figure 4: Tempo distribution for the eight musical genres of the "ballroom-dancer" database.

with the tempo information. The tempo we consider here has been manually entered for each track. We haven't considered the use of estimated tempo as [5] did.

In Figure 3, we represent the spectral rhythm pattern $\bar{Y}_n(\omega'_k)$ in the case of the product DFT/FM-ACF for all the songs belonging to each music genre category of the "ballroom dancer" database. Some characteristics of music genre appears immediately on this representation: ViennseWaltz and Waltz are the only genres having a component at $\omega' = 1/3$ (3/4 meter), but VienneseWaltz has no (a weak) component at $\omega' = 2$ (8^{th} note) while Walz has, Samba is the only genre having a component at $\omega' = 4$ (16^{th} note), Jive and QuickStep have no component at $\omega' = 2$. In Figure 4, we represent the tempo distribution for the eight musical genres of the database.

Classification algorithm We study the ability of each of the three feature sets to correctly classified the audio item of the "ballroom dancer" database into the 8 abovementioned classes. Each audio item is represented by a single 18-elements feature vector $Z(k)$. We consider three different classification algorithms: 1) the widely used C4.5. decision tree algorithm, 2) the Partial Decision Tree algorithm, 3) the Classification using regression methods. We have used the J48, PART, and ClassificationViaRegression implementations of Weka [10]. In order to allow to compare our results with the ones obtained by [4],

	J48	PART	ClassViaReg
DFT	75,64	73,78	**80,8**
ACF	69,34	70,34	76,64
DFT/FM-ACF	65,9	65,32	75,5
DFT + tempo	90,4	88,96	**90,4**
ACF + tempo	86,67	86,67	90,25
DFT/FM-ACF + tempo	86,38	86,24	90,25
tempo	77,79	77,36	77,93

Figure 5: Recognition rates obtained using various feature sets and classifiers (with and without tempo information)

classified as -->	C	J	O	R	S	T	VW	W
ChaChaCha	**87,4%**			4,5%	0,9%	7,2%		
Jive		**86,7%**	1,7%			6,7%	5,0%	
Quickstep		1,2%	**97,6%**	1,2%				
Rumba	2,0%			**79,6%**	2,0%	3,1%		13,3%
Samba	1,2%			7,0%	**89,5%**	1,2%		1,2%
Tango	3,5%			1,2%	1,2%	**94,2%**		
Viennese Waltz		3,1%			1,5%		**95,4%**	
Waltz		0,9%		4,5%				**94,5%**

Figure 6: Confusion matrix using DFT + tempo feature set and a Classification Via Regression algorithm

[5] and [6], we evaluate the performances using a 10-fold cross validation method.

Results: The recognition rates obtained using the three feature sets (with and without tempo information) with the various classification algorithms are indicated in Figure 5. In almost all cases, the best classifiers is the ClassificationViaRegression. The tempo alone achieves up to 78% correct recognition. Without the tempo information, the DFT is the best feature sets (81%), then the ACF (77%) and the product DFT/FM-ACF (75.5%). With the tempo information, all feature sets have very close recognition rates; however the DFT set performs slightly better (90.4%). In comparison, [4] report 90.1% recognition using a large set of features with the correct tempo (78.9% with the estimated tempo, 79.6% without the tempo), [6] report 50% using only the temporal rhythmic pattern, and 96% using this pattern, the whole set of features of Gouyon and the tempo. Note however that our representation does not use any acoustical feature. The **confusion matrix** is indicated in Figure 6. The larger confusion occurs between ChaChaCha/ Rumba/ Tango, Samba/ Rumba, Jive/ Tango/ VienneseWaltz and Rumba/ Waltz. These larger confusions can be explained either by their close tempi (see Figure 4) or their close spectral rhythm patterns (see Figure 3). In the opposite, in our study, the confusion between ViennseWaltz and Waltz remains low. **Best features:** In order to better understand, the discriminative power of each element k of $Z(k)$, we have applied an automatic feature selection algorithm (the Correlation Feature Selection of Weka) The first selected features are: $\frac{1}{3}$, $\frac{2}{3}$ (importance of ternary metrics), 1 (importance of the 4^{th} note) 2 (importance of the 8^{th} note) 3 (presence of 8^{th} note triplet), 3.75 (?) and 4 (importance of the 16th note). This corresponds to the intuition we get from Figure 3. Reducing $Z(k)$ to the 7 above-mentioned features, only slightly decreases the results: from 80.8% to 75.5% without tempo information, and from 90.4% to 89.54% with tempo information (using the DFT feature type and a ClassificationViaRegression algorithm).

4 CONCLUSION

In this paper, we have studied the use of three spectral patterns to represent the rhythm characteristics of an audio item. For a task of music genre classification, we have shown that the use of these simple spectral patterns allows to achieve a high recognition rate (close to the results obtained with more complex methods proposed so far). Among the three proposes spectral patterns, the use of a pattern derived from the DFT allows to achieve the highest recognition rate (90.4% with tempo, 81% without tempo). This result is surprising considering we though that the product DFT/FM-ACF would allow to better differentiate the various characteristics of rhythm. This is possibly due to the frequency mapping process of the FM-ACF, which decreases the overall frequency resolution. Future works will concentrate on evaluating the performances of this method when using the estimated tempo (instead of using the ground-truth tempo), and when applied to a larger set of music genre.

ACKNOWLEDGEMENTS

Part of this work was conducted in the context of the European IST project Semantic HIFI (http://shf.ircam.fr) [11].

REFERENCES

[1] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 10(5):293–302, 2002.

[2] J. Foote and S. Uchihashi. The beat spectrum: A new approach to rhythm analysis. In *ICME*. Pal Xerox FXPAL-PR-01-022, 2001.

[3] J. Paulus and A. Klapuri. Measuring the similarity of rhythmic patterns. In *ISMIR*, Paris, France, 2002.

[4] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer. Evaluating rhythmic descriptors for musical genre classification. In *AES 25th Int. Conf.*, 2004.

[5] F. Gouyon and S. Dixon. Dance music classification: a tempo-based approach. In *ISMIR*, 2004.

[6] S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *ISMIR*, Barcelona, Spain, 2004.

[7] G. Peeters. Time variable tempo detection and beat marking. In *ICMC*, Barcelona, Spain, 2005.

[8] P. Flandrin. *Time-Frequency/Time-Scale Analysis*. Academic Press, San Diego, California, 1999.

[9] Ballroom-Dancers.com.

[10] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Impl.* Morgan Kaufmann, San Fransisco, CA, 1999.

[11] H. Vinet. The semantic hifi project. In *ICMC*, Barcelona, Spain, 2005.