Evaluating rhythmic descriptors for musical genre classification

Fabien Gouyon[1], Simon Dixon[2], Elias Pampalk[2], and Gerhard Widmer[2,3]

[1] *Universitat Pompeu Fabra, Barcelona, Spain*

[2] *Austrian Research Institute for Artificial Intelligence, Vienna, Austria*

[3] *Department of Medical Cybernetics and Artificial Intelligence, Medical University of Vienna, Austria*

Correspondence should be addressed to Fabien Gouyon (`fgouyon@iua.upf.es`)

ABSTRACT

Organising or browsing music collections in a musically meaningful way calls for tagging the data in terms of e.g. rhythmic, melodic or harmonic aspects, among others. In some cases, such metadata can be extracted automatically from musical files; in others, a trained listener must extract it by hand. In this article, we consider a specific set of rhythmic descriptors for which we provide procedures of automatic extraction from audio signals. Evaluating the relevance of such descriptors is a difficult task that can easily become highly subjective. To avoid this pitfall, we assessed the relevance of these descriptors by measuring their rate of success in genre classification experiments. We conclude on the particular relevance of the tempo and a set of 15 MFCC-like descriptors.

1. INTRODUCTION

Currently, musical metadata and labels are largely produced manually, and such labels can serve, among many other things, to browse, or find one's way in musical databases. "Musically meaningful" labels greatly enhance such procedures.

Musical genre is a fundamental kind of metadata for browsing musical collections. Indeed, people often describe their musical tastes with respect to genre. Musical genre classification has received much attention from music record retailers and, recently, from audio and music researchers, especially in the Music Information Retrieval community [10, 1]. An important direction of research now relates to the definition of features of musical genres and their automatic extraction from various forms of musical data (audio, scores, MIDI, mp3, etc.).

Even if there is still room for disagreement on explicit definitions of musical genres [1], there is a pervasive belief that this notion has something to do with fundamental musical dimensions such as melody, instrumentation, harmony and rhythm. Rhythmic descriptors are therefore very valuable candidates for musical metadata. This article reports on definitions and evaluations of such descriptors.

1.1. Rhythm classification

A large amount of literature exists on the extraction of rhythmic descriptors from musical data, from symbolic data, audio or compressed audio. See [5] for an exhaustive review. There are indeed many ways to represent rhythm, from *low-level* signal-related quantities to more *abstract* concepts. Designing new descriptors is not a difficult task, but reliably extracting them from musical data in an automatic way is more difficult, even more so if they refer to cognitively relevant concepts. Even more problematic is the issue of their relevance in rhythm classification, similarity and retrieval tasks.

Many computer systems focus on the extraction of rhythmic descriptors defined by Western music theory. They focus on elements of the metrical structure, e.g. the tempo, the fastest pulse, the time signature, the quantized durations, the swing, the tempo variation or complete rhythmic transcriptions. Many articles point towards the usefulness of these metrical descriptors for rhythm classification and retrieval tasks. However, very few papers report on systematic assessments of their relevance for these tasks. In [3], Dixon *et al.* conclude that very few periodicities (the tempo, the measure and optionally others as the dotted quarter-note) seem sufficient to classify 8 rhythmic classes relatively well. The major source of error being in assigning the correct names ('quarter-note', 'measure', etc.) to detected periodicities.

Other researchers rather focus on descriptors more tightly linked to physical properties of the signal and whose musical meaning is less explicit. For instance, [10] and [8] respectively report on genre classification experiments and definitions of similarity distances using signal descriptors that, somehow, embed something about the rhythm. Also, [4] claims that all aspects of rhythm can be captured by a continuous periodicity representation and that such representation is sufficient for the retrieval of similar pieces of audio. However, this conclusion is based on the analysis of solely 15 musical excerpts (4 songs divided into several 10s chunks).

In this paper, following previous work by some of the authors [3], we assess the relevance of a set of rhythmic descriptors in automatic musical genre classification experiments. Genre-labelled musical data provides the necessary "ground-truth" for our experiments.

We acknowledge the fact that there actually exists no ground-truth with respect to genres [1]. However, some musical genres are rapidly recognisable by listeners, even with minimal musical training, and on the dancefloor, dancers do recognise instantly what dancing step fits best to the music they hear. Dancing having much to do with rhythm, it seems that ballroom dance music provides a relatively solid basis for our experiments. Here, we focus on excerpts of standard and Latin ballroom dance music, namely Jive, Quickstep, Tango, Waltz, Viennese Waltz, Cha Cha Cha, Samba and Rumba.

The organisation of the article is the following: we first give the details of the data and metadata used. Then we provide procedures of automatic extraction of rhythmic descriptors from audio signals. We then report on the method and results of rhythm classification experiments with these descriptors. We finally summarise and comment our results and propose directions for future work.

2. DATA AND ASSOCIATED METADATA

The musical database we use for training and testing contains excerpts from 698 pieces of music, around 30 seconds long. The audio quality of this data is quite low, it was originally fetched in real audio format, with a compression factor of almost 22 with respect to the common 44.1 kHz 16 bits mono WAV format. It was subsequently converted to WAV format for experiments. This data is publicly available on the world-wide web at the following URL:

http://www.ballroomdancers.com/Music/style.asp

For all those recordings, the musical genre is available. The data covers eight musical sub-genres of ballroom dance music:

- Jive, 60 instances
- Quickstep, 82 instances
- Tango, 86 instances
- Waltz, 111 instances
- Viennese Waltz, 65 instances
- Samba, 86 instances
- Cha Cha Cha, 111 instances
- Rumba, 97 instances

In addition, the tempo (in beats per minute, BPM) of each recording is also available. The minimum value is 60 BPM, the maximum 224 BPM.

3. DESCRIPTORS

We consider 73 descriptors, divided into three groups. Some of these descriptors have been computed using Matlab, the rest in C++, part of them with the free CLAM library.[1] All are implemented as open source software under the GNU license.

The first pool of descriptors (N=4) are widely used rhythmic concepts related to the metrical hierarchy:

[1] http://www.iua.upf.es/mtg/clam

- The ground truth tempo, in BPM, as provided with the data.

- BeatRoot tempo. As detailed in [2], BeatRoot's tempo induction stage yields several tempo hypotheses that are subsequently refined, beat by beat, and ranked in a tracking process. The final tempo is the mean of the winning agent's inter-beat intervals.

- "Naive" tempo. This is the highest peak of a periodicity representation (the inter-onset interval histogram, see below and [6] for more details). Note that this is one of BeatRoot's primary tempo hypotheses.

- The tick, or metrical level that coincides with all note onsets (i.e. the fastest level) [6].

For the second pool of descriptors, we consider 11 descriptors based on a first representation of signal periodicities, the "periodicity histogram" (PH) [8]. This representation (see Figure 1), loosely inspired by [10], is the collection in a histogram of the saliences of different pulses (from 40 BPM to 240 BPM) in successive chunks of signal (12s long, with overlap). In each chunk of signal, periodicities are computed via a comb filterbank [9]. Among relevant differences with previous works stands the fact that the audio data is first preprocessed by a psychoacoustic model, removing information in the audio signal which is not critical to our hearing sensation while retaining the important parts. Also, periodicity magnitudes are weighted w.r.t. their periods, emphasis being given to tempi around 120 BPM, the "preferred tempo" region.[2]
Descriptors are the following:

- The most salient periodicity: highest peak in the PH.

- The distinctiveness of the most salient periodicity. It is measured as the ratio between the highest peak and the second highest peak.

- The periodicity power. This is the sum of the energy in the PH.

[2]See [8] for details.

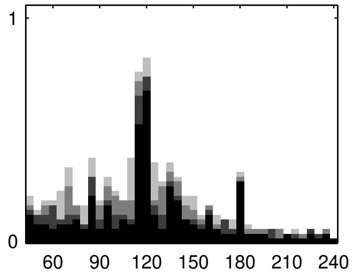

Fig. 1: Periodicity histogram of a Jive excerpt. The tempo is 176 BPM. Gray shadings tell us the number of analysis chunks for which a certain energy was exceeded. Note the effect of "preferred tempo" weighting.

- The periodic energy in the first three Bark bands. This is the same as the previous, but considering solely the energy in the 3 lowest frequency bands defined by the Bark scale, below 300 Hz.

- The PH centroid, defined as the tempo for which half of the PH energy is contained in lower tempi.

- Three measures of the percussiveness. The percussiveness is computed as the central tendency of the energy in diverse frequency bands, defined by the Bark scale, of the half-wave rectified, first-order difference filtered, waveform. We use three variations of this descriptor where the central tendency of the energy is computed in different ways: $mean(x)$, $mean(x > mean(x))$ and $median(x > median(x))$.

- Three measures of the percussiveness in low frequencies. This is similar as above but using only the energy in the 3 lowest Bark bands.

The third pool of descriptors (N=58) are quantities computed from a second representation of the signal periodicities, the inter-onset interval histogram (IOIH) [6]. This representation gives a measure of recurrence of the different inter-onset intervals present in the signal (not just successive onsets, but *any* pairs of onsets). Time intervals (in seconds) are drawn on the X-axis while (normalised) recurrences are drawn on the Y-axis (see Figure 2).

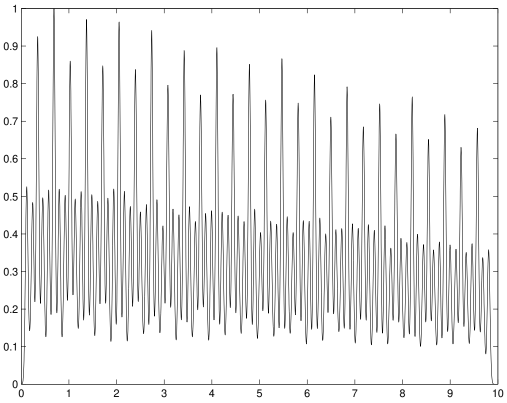

Fig. 2: IOI histogram of the same Jive excerpt as Figure 1. Recurrence vs time interval. The tempo is 176 BPM (around 350 ms), which corresponds to the third peak (not to the highest one). The second highest peak is the measure (44 MPM, around 1.4s).

Onsets are computed as in [7]. Then differences are computed and accumulated in a histogram which is then smoothed by a Gaussian window.[3] The IOIH is in many ways similar to the IOI clusters obtained in [3]. We therefore compute descriptors directly inspired by those detailed in [3]: selected prominent periods in the IOIH, together with their saliences.

- The saliences of 10 periodicities whose periods are the 10 first integer multiples of the tick. Note that solely the period *salience* is kept, not the period value. Therefore, those descriptors are independent of the tempo.

Then, inspired by the analogy between the IOIH and a spectral representation, we define 48 other descriptors as common "spectral" descriptors (distribution statistics and MFCCs), but computed on the IOIH, not on a spectrum. In the following $\{x_i\}_{i=1...N}$ are the IOIH samples.

- The mean of the IOIH magnitude distribution

$$\frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

[3]whose width is set to 300ms for IOIH descriptor computation and 150ms for tick computation.

- The geometric mean of the IOIH magnitude distribution

$$(\prod_i x_i)^{1/N} \tag{2}$$

- The IOIH total energy

$$\sum_{i=1}^{N} x_i^2 \tag{3}$$

- The IOIH centroid

$$\frac{\sum_{i=1}^{N} i * x_i}{\sum_{i=1}^{N} x_i} \tag{4}$$

- The IOIH flatness

$$\ln(gmean) - \ln(mean) \tag{5}$$

- The kurtosis of the IOIH magnitude distribution. It measures how outlier-prone a distribution is, i.e. its degree of peakedness [11].

$$\frac{\mu_4}{\mu_2^2} - 3 \tag{6}$$

where μ_2 and μ_4 are respectively the second and fourth central moments of the IOIH magnitude distribution [11].

- The IOIH "high-frequency content"

$$\sum_{i=1}^{N} i * x_i^2 \tag{7}$$

- The skewness of the IOIH magnitude distribution. This is the degree of asymmetry of a distribution [11]. A distribution spread out more to the left than to the right of the mean has a negative skewness. Perfect symmetry (e.g. a Gaussian distribution) results in a null skewness.

$$\frac{\mu_3}{\mu_2^{3/2}} \tag{8}$$

where μ_3 is the third central moment of the IOIH magnitude distribution [11].

- The first 40 coefficients of an analog to the Mel-Frequency Cepstral Coefficients (MFCCs).

MFCCs are widespread descriptors in speech research. The Cepstral representation has been shown to be of prime importance in this field, partly because of its ability to nicely separate the representation of voice excitation (the higher coefficients) from the subsequent filtering performed by the vocal tract (the lower coefficients).[4] Roughly, lower coefficients represent the spectral envelope (i.e. the formants) while higher ones represent finer details of the spectrum.
The CLAM implementation of the MFCCs is a porting of Malcolm Slaney's Matlab Toolbox,[5] to C++, with minor changes in the code. One way of computing the Mel-Frequency Cepstral representation of a time signal is as follows:

- Short-time windowing
- Fourier transform - keep solely the magnitude spectrum
- Projection of the frequency axis from linear scale to the Mel scale, of lower dimensionality (usually by means of a filterbank)
- Magnitude logarithm computation
- Inverse Fourier transform

In our case, we follow the same steps, apart from the first two that are replaced by the computation of the IOIH.

4. METHOD

Before giving the detail of any classification experiment, we must note that the baseline for classification is 15.9% (classification rate when always guessing the most probable class). This value should be kept in mind when evaluating the goodness of any classifier. Experiments have been conducted using the free software Weka,[6] we refer to [12] for any details regarding algorithms mentioned below. All classification accuracies reported below are computed as 10-fold cross-validations: 10 subsets containing 90% randomly selected samples are selected for learning and the remaining 10% are used for testing, final percentages are averages over those 10 runs.

[4] http://mi.eng.cam.ac.uk/~ajr/SA95/
[5] http://www.slaney.org/malcolm/pubs.html
[6] http://www.cs.waikato.ac.nz/ml/weka/

Our chief objective is the relevance ranking of the descriptors with respect to rhythm classification more than the design of a reliable, generalizable and ready-to-use classifier. Therefore, we used few classifiers in the following experiments. We rather focused on an evaluation of the descriptors with respect to the same classifier, most of the times a Nearest Neighbour scheme (1-NN).

4.1. Attribute selection

The number of input attributes to the classification algorithm must be reduced for three reasons: getting simpler models, improving prediction accuracy (particularly with some classifiers which suffer from the "curse of dimensionality", such as k-NN), and to get more insight into which aspects are relevant.

In some cases, attribute selection can be driven by our understanding of the attribute meanings and our intuitions regarding their relevances. In addition we used the following automatic attribute selection methods:

- Evaluation of the attributes on an *individual* basis (use of the Ranker search method), associated to two different attribute evaluation methods, ReliefF (method1) and symmetrical uncertainty (method2). This is done 10 times on different samples (of size 90%) drawn from the dataset. The final ranks of the attributes are the averages of the 10 runs.
- Evaluation of attribute *subsets*. We selected the forward search method to explore the attribute space. Different subset evaluators have been tried:
  - Correlation-based (Cfs), performed 10 times on different samples (of size 90%) drawn from the dataset, the final attribute ranks being the averages of the 10 runs. (method3)
  - Wrapping around a 1-NN classifier whose accuracy is determined by 5-fold cross-validations. This is done 10 times on different samples (of size 90%) drawn from the dataset. The final attribute ranks being the averages of the 10 runs. (method4)

- Wrapping around a 1-NN classifier whose accuracy is determined by 10-fold cross-validations. This is done just 1 time on the whole dataset. (method5)

5. EXPERIMENTS AND RESULTS

5.1. Relevance of the tempo

There is a common belief that tempo is of prime relevance for classifying musical pieces. We tested this assumption in the following experiments.

5.1.1. Correct tempo

As previously commented, the source of audio data also provides ground-truth tempo values for each excerpt. A simple Nearest Neighbour classifier (k-NN) using solely this descriptor reaches a classification accuracy of **82.3%** (with k=1). A C4.5 decision tree achieves 78.6%. This last result was obtained with a special tweaking of the algorithm: forcing a relatively high number of instances per leaf (20 instead of default value 2), which results in smaller trees, with fewer leaves and guarantees good generalization of the result. The number of leaves is 9. Each class corresponds to a leaf, except one (Rumba) which corresponds to two leaves. In sum, this technique highlights a clear ordering of classes w.r.t. tempi, from slow to fast:

$\text{tempo} < 91 \Rightarrow$	Waltz
$91 < \text{tempo} < 96 \Rightarrow$	Rumba
$96 < \text{tempo} < 102 \Rightarrow$	Samba
$102 < \text{tempo} < 104 \Rightarrow$	Rumba
$104 < \text{tempo} < 124 \Rightarrow$	ChaChaCha
$124 < \text{tempo} < 141 \Rightarrow$	Tango
$141 < \text{tempo} < 176 \Rightarrow$	Jive
$176 < \text{tempo} < 180 \Rightarrow$	VienneseWaltz
$\text{tempo} > 180 \Rightarrow$	Quickstep

Further analysis of the results showed that strongest confusions are between Rumba and Samba or Tango and Cha Cha Cha.

The next paragraph shows that such performances are not achievable when using computed tempo values instead of the manually given tempo.

5.1.2. Computed tempo

Using solely BeatRoot tempo, a 1-NN classifier yields **51,7%** correct classification, against 41.2% for the "naive" tempo computation. With C4.5, with the same parameter set as above, the results are, in the same ordering, 52.5% and 42.5%.

As shown in Figure 3, the tempo induction algorithms detailed in [2] and [6] make systematic errors by confusing metrical levels. The former is more accurate than the latter as it could be expected. This explains why BeatRoot tempo is a better descriptor for classification that the "naive" tempo.

However, accuracy levels are far from those with correct tempi. First, classification rates are much lower. But more importantly, for the 2 algorithms, decision trees yield too many leaves (between 13 to 17). This means that they divide the tempo axis in small clusters. We lost the nice "one tempo per class" scheme because of errors in the choice of metrical level common proper to tempo induction algorithms. Note that those are errors common to all state-of-the-art algorithms and no solution to this problem is to be expected soon (indeed, it is a very "natural" error to tap the beat at half or twice its speed).

Focusing on the fastest pulse (the tick) instead of the perceptual beat, correct classification rates are comparable: 50.4% with a 1-NN classifier and 51.8% with C4.5 (15 leaves).

5.2. PH descriptors

With the 11 PH descriptors, a 1-NN scheme yields 52.8% correct classification. The classification rate

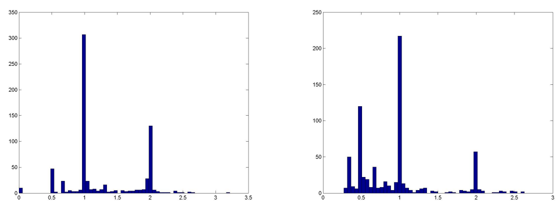

Fig. 3: Tempo extraction performance of two algorithms (respectively [2] and [6]). The plots show computed tempi divided by correct tempo, X=1 means "exact-match", the highest peak for both algorithms. Other high peaks are at X=1/2, 2, 1/3 and 2/3.

can be kept around the same value (slightly higher, **56.7%**) when discarding 6 descriptors, by inspection of the results of the 5 feature selection methods, and keeping solely:

- The most salient periodicity
- The distinctiveness of the most salient periodicity
- The periodicity power
- The PH centroid
- The first measure of the percussiveness in low frequencies

5.3. Periodicity saliences

We refer here to the magnitudes of the IOIH peaks whose periods are the ten first integer multiples of the tick.

When using the 10 IOIH peak ampitudes together with the correct tempo, we reach a 75.5% of correct classification with a 1-NN classifier. This is worse than with the correct tempo alone. However, it does not mean that these descriptors are not relevant. Indeed, when removing the tempo from the attribute set, keeping therefore solely descriptors that are *independent of the tempo*, we reach a **51.2%** of correct classification. This is only slightly lower than the performance when associated to the computed tempo (BeatRoot version), 54.3%, and similar to the performance of the computed tempo alone.

A reason for this phenomenon is that instances misclassified when focusing on the tempo alone are *also* misclassified when considering solely the IOIH peaks. For instance, with the correct tempo alone, an inspection of the confusion matrix reveals that the common misclassifications are Rumba classified as Samba (17 out of 97). Similarly, solely 37% of the Rumba are correctly classified with the IOIH peaks.

So, on the one hand, it is clear that we should account somehow for the pace of the music and on the other, computed tempo associated to IOIH peaks does not show clear advantages. However an association IOIH peaks - tick (which performed relatively well when considered alone) performs better: **65.1%** of correct classification with a 1-NN classifier.

5.4. Other IOIH descriptors

Let us consider the first 8 distribution statistics (i.e. not the MFCC-like). Using them all yields 46.1% classification accuracy with a 1-NN classifier. Selecting solely 3 by inspecting results of the 5 automatic attribute selection methods yields a slight improvement: **48.7%**. “Winning” descriptors are:

- The kurtosis
- The skewness
- The high-frequency content

Let us now consider the MFCC-like descriptors. Also with 1-NN classification, the whole pool (i.e. 40 descriptors) yields **79.6%** accuracy. Here also, the dimensionality can be reduced automatically (we did not consider method4, being too computationally expensive with 40 descriptors). A very similar classification accuracy can be reached (79%) when selecting 15 coefficients: 1, 2, 3, 6, 7, 8, 10, 11, 15, 16, 19, 24, 25, 26 and 28.

5.4.1. On the meaning of IOIH MFCC-like descriptors

When dealing with speech signals, it has been shown that most of the relevant information occurs near the origin of the cepstral representation and in a few peaks higher up the cepstrum,[7] these peaks corresponding to multiples of the pitch. Hence the focusing on the first MFCCs (less than 20), providing a compact representation of the spectral envelope while discarding the fine detail pitch information. This is especially true in speech recognition tasks where researchers precisely seek pitch-independent descriptors.

When dealing with music signals, and when replacing the Fourier transform by an ad-hoc transformation (the IOI histogram), it is less clear that higher coefficients should be discarded. In our case, higher coefficients provide a representation of finer detail of the IOIH peaks, that is, a closer look at the harmonic nature of this periodicity representation, its “pitch.”[8] Therefore, higher coefficients seem to be somehow related to the pace of the piece at hand.

[7] http://mi.eng.cam.ac.uk/~ajr/SA95/

[8] Note that the tick is precisely computed as the “gap of the IOIH harmonic series” [6].

On the other hand, lower coefficients represent the global envelope of the IOIH, which would be the "spectral envelope" of a proper spectrum. They seem to represent in some way the global structure of the IOIH.[9] In our understanding, they encode some aspects of the metrical hierarchy. Independently of the tempo.

5.5. All in a bag

Consider now the subset of 35 descriptors that showed to be relevant in the previous experiments (the tick, the tempo, 5 PH descriptors, 10 periodicity saliences, 3 IOIH distribution statistics and 15 MFCC-like) and let us keep on with the selection of the most relevant.

Associating the 15 MFCC-like descriptors with the correct tempo, the accuracy reaches **90.1%**, the best result until now. However, we must stress here that if we use the computed tempo or the tick as representative of the musical pace instead of the correct tempo, the performance decreases: 78.9% and 77,2% respectively, which is slightly worse than the MFCC-like descriptors alone. In fact, running the feature selection methods aforementioned and using intuitions regarding which descriptors "should make it", we did not find any feature subset[10] that would present *significant* improvements over the 15 MFCC-like descriptors alone.

However, an interesting result is that we can lower the dimensionality to 9 descriptors that do *not account for the tempo* (periodicity power and MFCC-like coefficients 1, 6, 7, 8, 10, 15, 16 and 24), and keep a comparable accuracy as with the 15 MFCC-like (77.3%).

6. SUMMARY, DISCUSSION AND FUTURE WORK

The tempo is a very relevant feature for genre classification. Considering solely this feature gives very good results (over 80% accuracy). However, tempo values given by beat induction algorithms are much less useful than the correct tempo, assigned manually. This is due to common errors in metrical level. This is in accordance with previous work [3] that stressed on the one hand the relevance of metrical levels for rhythm classification and the other the difficulty of deriving them unambiguously from a periodicity representation. Indeed, any metrical level corresponds to a peak in the IOIH,[11] but the contrary is not true, not all peaks are part of the metrical hierarchy. That is, not all periodicities are metrical levels.
Representing the musical pace by the tick is also relevant as it performs similarly as the computed tempo.
Another conclusion is that in order to compute a tempo value that would best resemble a human perception of the musical pace, it is better to consider a whole tracking process in addition to the periodicity induction (as BeatRoot does) rather than rely solely on the induction (as the "naive" method does). In other words, the correct tempo might not be the highest peak in a periodicity representation, but rather the one that propagates better onto the whole data.

Other descriptors of periodicity representations, with weaker or less explicit musical meanings, are also relevant. Several representations are possible (e.g. IOIH, PH) and they can be parameterised in different manner (e.g. peaks, distribution statistics, MFCC-like descriptors). With few low-level descriptors (e.g. 5 PH descriptors, 10 IOIH peaks, 3 IOIH descriptors), classification accuracies are encouraging.
Among the set of periodicity representation descriptors, the subset that provides the *best trade-off accuracy-dimensionality* is a set of 15 Mel-Frequency Cepstrum Coefficients computed on the IOIH (79% accuracy with only 15 descriptors).
In our understanding, higher coefficients are related to the musical pace and lower coefficients to the metrical hierarchy.

Associating the correct tempo with the 15 MFCC-like descriptors gives the best result: 90% accuracy.

Without the correct tempo (as in real life), we encourage the selection of the 15 aforementioned MFCC-like descriptors, or a subset of 9 that does not damage the accuracy too much.

We illustrated the fact that one can not simply increase the number of descriptors and wish that this

[9] For instance, excerpts whose periodicities have very similar saliences, as e.g. many Cha Cha Cha, have a flat envelope.

[10] that would not contain the correct tempo

[11] or is one of the IOI clusters of [3]

will result in an improvement of the classification accuracy, even if all the descriptors have shown some discriminative power. The descriptors introduced above are indeed representative of the rhythm but, in our understanding, they represent rhythm only *partly*. Some aspect of rhythm, that would permit to separate more accurately those 8 classes, is lacking.

Future work is related to the incorporation of new descriptors (e.g. the swing ratio, the time signature and the syncopation factor) and the evaluation of their relevance on a larger amount of data. We also strongly believe that *time-varying* descriptors should be accounted for, as well as some representation of typical temporal *patterns*.

The design of a rhythmic distance, accounting for the aforementioned descriptors, is another direction of future work. This would permit to make an important step: organising *any* song collection with respect to rhythm, not just songs which we already know belong to a restricted set of classes.

7. ACKNOWLEDGMENTS

This research was partly supported by the project Y99-INF, sponsored by the Austrian Federal Ministry of Education, Science and Culture (BMBWK) in the form of a START research prize. The Austrian Research Institute for Artificial Intelligence acknowledges the basic financial support from the BMBWK and the Ministry for Transport, Innovation and Technology (BMVIT).

This work reported in this paper was also partially funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). More information can be found at the project website http://www.semanticaudio.org.

Thanks also to Perfecto Herrera and Pedro Cano for general comments, to Eloi Batlle for comments on MFCCs and to all the people in the MTG who contributed to that part of the CLAM code used to carry on these experiments (Gilles Peterschmitt, Julien Ricard, Guenter Geiger, David Garcia, Lars Fabig, etc.).

8. REFERENCES

[1] J.J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," Journal of New Music Research, 32(1), 2003.

[2] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," Journal of New Music Research, 30 (1), 2001.

[3] S. Dixon, E. Pampalk and G. Widmer, "Classification of dance music by periodicity patterns," Proceedings of the International Conference on Music Information Retrieval, 2003.

[4] J. Foote, M. Cooper and U. Nam, "Audio retrieval by rhythmic similarity," Proceedings of the International Conference on Music Information Retrieval, 2002.

[5] F. Gouyon, S. Dixon, "A review of automatic rhythm description systems," submitted.

[6] F. Gouyon, P. Herrera and P. Cano, "Pulse-dependent analyses of percussive music," Proceedings of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, 2002.

[7] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1999.

[8] E. Pampalk, S. Dixon, and G. Widmer. "Exploring music collections by browsing different views," Proceedings of the International Conference on Music Information Retrieval, 2003.

[9] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," J. Acoust. Soc. Am. 103(1), 1998.

[10] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Transactions on Speech and Audio Processing, 10(5), 2002.

[11] E. Weisstein, from MathWorld–A Wolfram web resource. http://mathworld.wolfram.com/topics/Moments.html

[12] I. Witten and E. Frank, "Data Mining: Practical machine learning tools with Java implementations," Morgan Kaufmann, San Francisco, 2000.