ON RHYTHM AND GENERAL MUSIC SIMILARITY

Tim Pohle[1], Dominik Schnitzer[1,2], Markus Schedl[1], Peter Knees[1] and Gerhard Widmer[1,2]
[1]Department of Computational Perception, Johannes Kepler University, Linz, Austria
[2]Austrian Research Institute for Artificial Intelligence (OFAI), Wien, Austria
music@jku.at

ABSTRACT

The contribution of this paper is threefold:

First, we propose modifications to Fluctuation Patterns [14]. The resulting descriptors are evaluated in the task of rhythm similarity computation on the "Ballroom Dancers" collection.

Second, we show that by combining these rhythmic descriptors with a timbral component, results for rhythm similarity computation are improved beyond the level obtained when using the rhythm descriptor component alone.

Third, we present one "unified" algorithm with fixed parameter set. This algorithm is evaluated on three different music collections. We conclude from these evaluations that the computed similarities reflect relevant aspects both of rhythm similarity *and* of general music similarity. The performance can be improved by tuning parameters of the "unified" algorithm to the specific task (rhythm similarity / general music similarity) and the specific collection, respectively.

1 INTRODUCTION

Many of the rhythm descriptors proposed so far eventually reduce the rhythm to a representation that discards information about which frequency band the rhythmic feature originates from. We begin this paper by asking: *"Can the performance of rhythm descriptors be improved by adding frequency information?"* To this end, we follow two directions. First, we propose and evaluate descriptors that retain information about the frequency range in which a given rhythm feature (more precise: periodicity strength) was measured. Related work in this direction includes [10]. Second, we add frequency information in the form of a "timbral" component (cf. [3]).

The paper is organized as follows. In Section 2, we suggest a number of modifications to Fluctuation Patterns (FPs) [14]. Relative to our evaluation setting, the modified variant seems to capture rhythmic similarity better than the unmodified algorithm. In Section 3, we go on by adding frequency information to the proposed rhythm descriptors in the form of a "timbral" component, and find that in our evaluation setting, rhythm similarity computation is improved further this way. We consider this finding as complementary to the practice of using rhythm descriptors to improve the performance of (general) music similarity measures (e.g., [14]). Based on this observation, we design an algorithm that seems to perform well *both* in the task of rhythm similarity *and* in the task of general music similarity computation (Section 4). In our evaluation setting, this combined algorithm outperforms approaches that are specifically designed for the respective tasks.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2009 International Society for Music Information Retrieval.

2 GETTING THE RHYTHM

This section is dedicated to rhythm descriptors and their evaluation on the Ballroom Dancers collection.

2.1 Rhythm Descriptors

Below, the rhythm descriptors evaluated in this paper are described. These are the well-known Fluctuation Patterns, and our proposed extensions *Onset Patterns* (OPs) and *OnsetCoefficients* (OCs).

2.1.1 Fluctuation Patterns (FPs)

Fluctuation Patterns (FPs) [14] measure periodicities of the loudness in various frequency bands, considering a number of psychoacoustic findings. We use the implementation of the MA Toolbox [1] with the proposed parameter set, so that the frequency bands correspond to 20 critical bands. Details about the computation are given e.g. in [14]. An evaluation of the importance of the various psychoacoustic processing steps in FP calculation is given in [10].

2.1.2 Onset Patterns (OPs)

We suggest a number of changes to FPs (cf. [4, 17, 18]). To this end, a number of preliminary experiments was conducted. The most important changes to FPs are listed here, before the points are discussed in detail:

- Reduce the signal to the parts of increasing amplitude (i.e., likely onsets).
- Use semitone bands to detect onsets instead of fewer critical bands.
- Use Hanning window and zero padding before detecting periodicities with FFT.

[1] http://www.pampalk.at/ma/

- Represent periodicity in log scale instead of linear scale.

We only consider onsets (or increasing amplitudes) in a given frequency band. To detect such onsets, we use a cent-scale representation of the spectrum with 85 bands of 103.6 cent width, with frames being 15.5 ms apart. On each of these bands, an *unsharp-mask* like effect is applied by subtracting from each value the mean of the values over the last 0.25 sec in this frequency band, and half-wave rectifying the result. This aims to detect also slow-attack instrument onsets in melodies that have notes with only one (or few) semitones apart. Subsequently, values are transformed by taking the logarithm, and reducing the number of frequency bands from 85 to 38 which is closer to the number of critical bands.

As in the computation of FPs, segments of frames are analyzed for periodicities. We use segments of 2.63 sec length with a superimposed Hanning window, zero-padded to six seconds. Adjacent segments are 0.25 sec apart. Each of these segments is analyzed for periodicities in the range from $T_0 = 1.5$ sec up to about 13.3 Hz (40 to about 800 bpm), separately in each of the 38 frequency bands. A crucial point in this transformation is that we do not represent periodicities on a linear scale (as in FPs), but rather we use a log-representation. Thus, after taking the FFT on the six seconds of a given frequency band, a log filterbank is applied to represent the selected periodicity range in 25 log-scaled bins. In this representation, periodicity (measured in Hz) is doubled every 5.8 bins (i.e., going 6 bins to the right means measuring a periodicity about twice as fast). By using this log scale, all activations in an OP are shifted by the same amount in the x-direction when two pieces have the same onset structure but different tempi. While this representation is not blurred (as done in the computation of FPs), the applied techniques induce a smearing in the lower periodicity range (cf. Figure 1). After a segment is computed, each of the 25 periodicities is normalized to have the same response to a broadband noise modulated by a sine with the given periodicity. This is done to eliminate the filter effect of the onset detection step and the transformation to logarithmic scale.

To arrive at a description of an entire song, the values over all segments are combined by taking the mean of each value over all segments. We call the resulting representation of size $38 \cdot 25$ *Onset Patterns* (OPs). In this paper, the distance between OPs is calculated by taking the Euclidean distance between the OPs considered as column vectors.

2.1.3 *OnsetCoefficients (OCs)*

OnsetCoefficients are obtained from all OP segments of a song by applying the two-dimensional discrete cosine transformation (DCT) on each OP segment, and discarding higher-order coefficients in each dimension. The DCT leads to a certain abstraction from the actual tempo (cf. [5, 18]) *and* from the frequency spectrum (like in MFCCs). This is motivated by the notion that slightly changing rhythm and sounds does not have a big impact on the perceived characteristic of a rhythm, while the same rhythm played with a drastically different tempo may have a different perceived characteristic. For example, one can imagine that a slow and laid-back drum loop, used in a Drum'n'Bass track played back two or three times as fast, is perceived as cheerful.

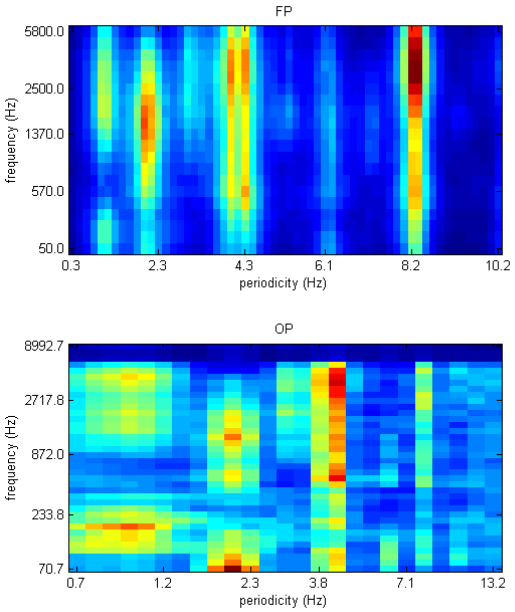

Figure 1. FP and OP of the same song. Doubling of periodicity appears evenly spaced in the OP. A bass drum plays at regular rate of about 2 Hz. The piece has a tap-along tempo of about 4 Hz, while the measured periodicities at about 8 Hz are likely caused by offbeats in between taps.

The number of DCT coefficients kept in each dimension (periodicity / frequency) is an important parameter. The selected coefficients are stacked into a vector. For example, keeping coefficients 0 to 7 in the periodicity dimension, and coefficients 0 to 2 in the frequency dimension yields a vector of length $8 \cdot 3 = 24$. We abbreviate this selection as 7×2. Based on the vectors for all segments, the mean and full covariance matrix (i.e, a single Gaussian) is calculated, which is the OC feature data for a song.

The OC distance D between two Songs (i.e., Gaussians) X and Y is calculated by the Jensen-Shannon (JS) divergence (cf. [11]).

$$D(X,Y) = H(M) - \frac{H(X) + H(Y)}{2} \qquad (1)$$

where H denotes the entropy, and M is the Gaussian resulting from merging X and Y. We calculate the merged Gaussian following [20]. We use the square root of this distance.

2.2 Setup for Rhythm Experiments

We evaluate the rhythm descriptors on the ballroom dance music set [2] previously used by other authors, e.g. [5, 4, 2,

[2] data from ballroomdancers.com

15, 7] and for the ISMIR'04 Dance Music Classification Contest [3]. This set consists of 698 tracks assigned to 8 different dance music styles ("genres"). The classification baseline is 15.9%.

The purpose of the descriptors discussed above is to measure *rhythmic similarity*. For evaluation, we assume that tracks that are in the same class have a similar rhythm. To facilitate comparison to previous work [5, 4], we use a 1-nearest-neighbor (1NN) stratified 10-Fold cross validation (averaged over 32 runs) in spite of a certain variance induced by the random selection of folds. We assume that the only information that is available is the audio signal. Using 1NN 10fold cross validation, [5] report up to 79.6% accuracy.

When using more sophisticated classification algorithms (and other features), higher accuracies are obtained. For example, [2] report a classification accuracy of up to 82% using only automatically computed features (i.e., without using correct tempo annotation or manually corrected first bar annotations). The highest classification accuracy we are aware of is 86.9%, obtained by kNN classification [7].

The mentioned accuracies are obtained when the audio signal is the only data source made available to the algorithms. It has to be noted that the algorithms yield higher accuracies when also the *correct* tempo annotation is given as feature data. In this case (which is not considered in this paper), an accuracy of 95.1% (or 96.0% when also human-corrected bar annotations are used [2]) have been obtained.

2.3 Results for Rhythm-Only Descriptors

FPs as implemented in the MA toolbox, compared by Euclidean distance, yield an accuracy of **75.0%**. OPs compared with Euclidean distance yield **86.7%**. The results for various settings of using only OnsetCoefficients for similarity estimation are shown in Figure 2. It can be seen that the highest values are obtained when keeping more than 16 coefficients in the periodicity dimension and when only keeping the 0th coefficient in the frequency dimension (which corresponds to averaging over all frequencies). In this range, values increase when including more periodicity coefficients, which seems consistent with the findings in [5]. In this range, we obtain an average value of **87.7%** [4].

3 ADDING "TIMBRE" INFORMATION

To examine how the discussed rhythmic descriptors can be used in conjunction with "bag of frames" audio similarity measures, we combine them with a "timbral" audio similarity measure. The used frame-based features are the well-known MFCCs (coefficients 0..15), Spectral Contrast Coefficients [9] (using the 2N approach [1], coefficients 0..15), and the descriptors *Harmonicness* and *Attackness*. The latter two describe the amount of harmonic and percussive elements (cf. [13]) in a cent-scaled spectrogram with frequency bands being 66 cent and frames being 46 ms apart. Percussive elements are detected by applying a 5×5 filter with the kernel (-0.14, -0.06, 0.2, 0, 0) replicated over five rows. The analogous filter to detect harmonic elements has the form $(-0.09, -0.01, 0.2, -0.01, -0.09)^T$, replicated over five columns. The Harmonicness value for a frame is the sum of the half-wave rectified responses of this filter centered at the frequency bins of the considered frame. The frame's Attackness value is calculated the same way but using the filter for percussive elements. Altogether, these are 34 descriptor values for a frame, which are combined over a song by taking their mean and full covariance matrix. Two songs are compared by taking the Jensen-Shannon divergence as described above.

We combine the discussed rhythm descriptors with this timbral component by simply summing up the two distance values (i.e., timbral and rhythm component are weighted $1:1$). For comparison, e.g., in the G1C algorithm [14], FP based features are weighted with 30%, and a MFCC component is weighted with 70%. Our weighting decision is not based on systematic evaluations but rather it is mainly based on impressions gained from non-representative listening experiments. To bring the two distances (rhythm based and timbre based) to a comparable magnitude, for each song the distances of this song to all other songs in the collection are normalized by mean removal and division by standard deviation [5]. Subsequently, the distances are symmetrized by summing up the distances between each pair of songs in both directions. This preprocessing step is done for each component (timbral and rhythm) independently before summing them up.

3.1 Combination Experiment

We repeat the experiment shown in Figure 2, but this time combining the rhythm descriptors with the timbral component as described. The 1NN 10fold cross validation accuracy is 54.0% when considering only the timbral component, 79.4% in combination with FPs, and 87.1% with OPs. From the results in Figure 3, it can be seen that classification results are improved when combining OCs with the timbral component. This time, average results of 90.2% are obtained over the parameter range discussed above (compared to 87.7% in the the first experiment, Figure 2). The highest obtained 1NN accuracy is 91.3%.

Results are summarized in Table 1. The results for the combined method are above the values obtained for each component (rhythm and timbre) alone. We think this is an indication that rhythm similarity computations can be improved by including timbre information. This is in line with [19] who reason that tempo can be detected better when considering timbre information. In a way, this is complementary to previous approaches where descriptors of rhythmical properties were added to timbre descriptors in order to improve music similarity computations (e.g. the

[3] http://mtg.upf.edu/ismir2004/contest/rhythmContest/

[4] We take the average rather than the maximum value as an indicator due to variances introduced by 10fold CV.

[5] This is done once before splitting up training and test sets for classification. No class labels are used in this step. We expect the impact of determining the normalization factors only on the respective (stratified) training set to be negligible.

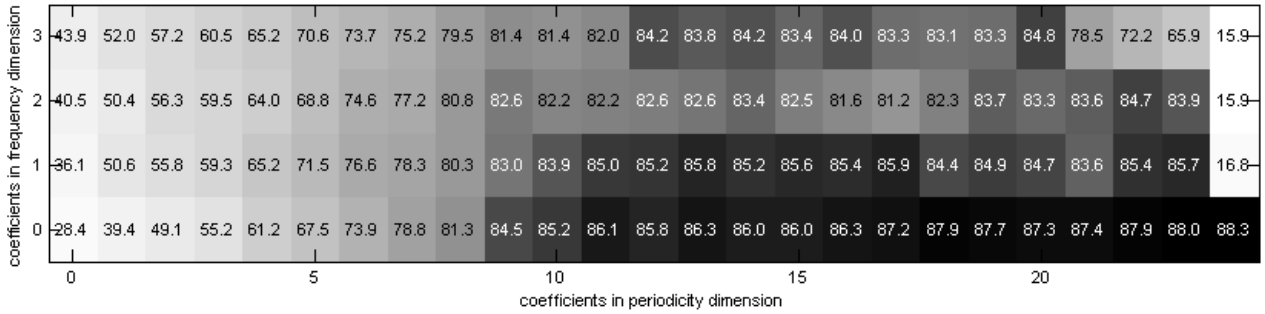

Figure 2. Dance genre classification based on OnsetCoefficients; distances calculated according to Equation 1. 1NN 10fold CV accuracies obtained on ballroom dataset when including coefficients 0 up to the given number in the respective dimension. For example, including coefficients 0..17 in the periodicity dimension and coefficients 0..1 in frequency dimension (resulting in $18 \cdot 2 = 36$ dimensional feature data) yields an accuracy of 85.9%. Low results at right border are caused by numerical instabilites when calculating the determinant during entropy computation. For better visibility, gray shades indicate ranks instead of actual values.

Algorithm	1NN
Baseline	15.9%
FP	75.0%
OP	86.7%
OC	up to around 87.7%
Timbre	54.0%
Timbre+FP	79.4%
Timbre+OP	87.1%
Timbre+OC	up to around 90.2%

Table 1. Ballroom dataset: 10fold CV accuracies obtained by the evaluated methods. The methods below the line are combined by distance normalization and addition.

G1C algorithm [14]). This duality leads to the experiments presented next.

4 THE "UNIFIED" ALGORITHM

Encouraged by the experiments presented in the previous section, we examine the performance of this algorithm not only in the task of *rhythm* similarity computation, but also in the task of general music similarity. Our aim is to find a selection of OCs that perform well in both tasks, which eventually leads to a "unified" music retrieval algorithm that reflects both rhythm and timbre similarity.

4.1 Data Sets

Music similarity experiments are performed on the set from the ISMIR'04 genre classification contest (ISMIR'04) [6], and on the "Homburg" data set (HOMBURG) [8]. Like the ballroom set, these collections are available to the research community, which facilitates reproduction of experiments and gives a benchmark for comparing different algorithms. There are two variants of the ISMIR'04 collection. The first is the "training" set which consists of 729 tracks from six genres. The second consists of all the tracks in the "training" and "development" sets, which are 1458 tracks from six genres. We use the central two minutes from each track. The HOMBURG set consists of 1886 excerpts of 10 seconds length.

[6] http://ismir2004.ismir.net/genre_contest/index.htm

4.2 Combination Experiment

In this section, we conduct a similar experiment as in Section 3.1 on the ISMIR'04 *training* collection. The aim is to evaluate the impact of OCs on the performance in *general* music similarity computation (i.e., not limited to rhythm similarity). The results from these experiments are used to create the "unified" algorithm, which will then be evaluated on all three collections (including the HOMBURG collection).

Following previous work [1, 14], we take genre classification accuracy as an indicator of the algorithm's ability to find similar sounding music. We use the same evaluation methodology as before. The timbre component alone yields 83.8%. Combining it with FPs as described, accuracy drops to 83.6%. Using OPs instead, accuracy increases to 85.2%. With OCs, accuracy can be improved up to 87.8% in the parameter range shown in Figure 4. This figure shows an outlier for 19×0 OCs, for which unfortunately we did not find an obvious explanation such as outliers in the distance matrix or numerical instabilities. Comparing Figures 3 and 4, it seems that a good tradeoff between the two collections is found when using 16×1 OCs. This selection yields $17 \cdot 2 = 34$-dimensional feature data, i.e., the rhythm feature data consists of a mean vector of length 34 and a covariance matrix of size $34^2 = 1156$.

4.3 Final Evaluation and Optimization

In Table 2, 10fold CV results obtained with this setting are listed. For comparison to previous work, also the highest classification accuracies obtained so far that we are aware of are listed. These accuracies refer to methods only using audio descriptors without additional human-annotated clues. On all three collections, the results of the "unified" algorithm are above these previously reported results.

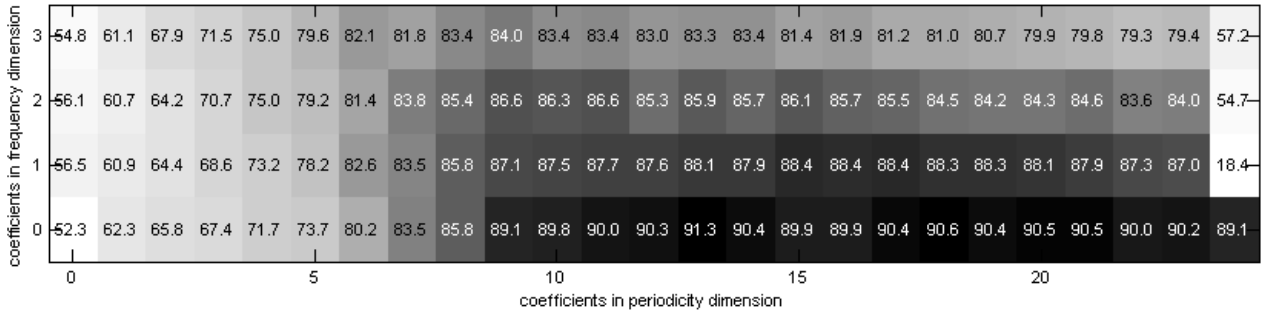

Figure 3. Combination of OCs with timbral component on the ballroom dancers collection, 1NN 10fold cross validation.

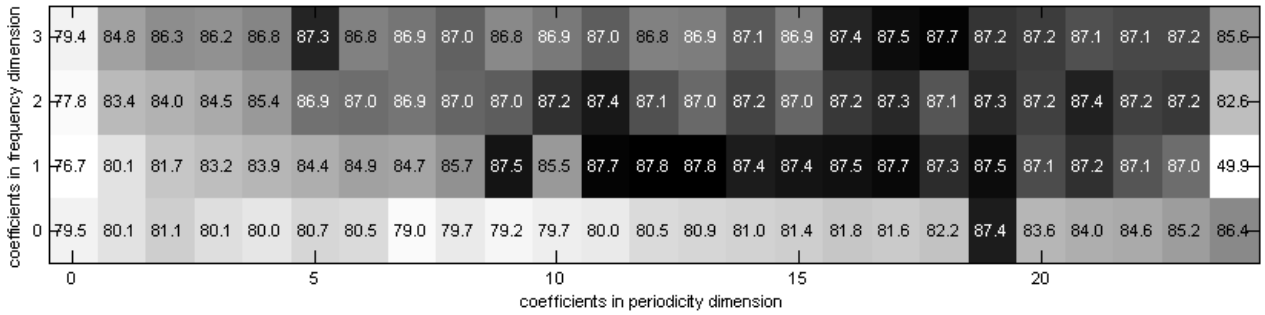

Figure 4. Combination of OCs with timbral component, ISMIR'04 training collection.

Collection	1NN	highest kNN (obtained at k)	Literature
Ballroom	88.4%	**89.2%** (k=5)	86.9% [7]
ISMIR'04 train	87.6%	**87.6%** (k=1)	84.0% [16]
ISMIR'04 1458	90.4%	**90.4%** (k=1)	83.5% [6]
HOMBURG	50.8%	**57.0%** (k=10)	55% [12]

Table 2. Accuracies obtained by the "unified" algorithm on the various collections.

While these results show that our "unified" algorithm outperforms the respective specialized approaches, we observe that when tuning to the particular collections, our techniques can be used to obtain even higher accuracies. For these experiments, we use leave-one-out evaluation for two reasons. First, doing 10fold cross validation (and repeating it several times for averaging) has a clearly longer runtime, as we evaluate a fixed matrix of pairwise distances. Second, in the 10fold cross validation experiments, we observe a certain variance between repeated experiments.

Our non-exhaustive tuning experiments indicate that even the normalization step used to combine two measures (Section 3) alone in some cases increases accuracy. On the Ballroom Dancers collection, a 3NN accuracy of 91.8% is obtained when including normalised OCs up to 24×0. Using only the normalised timbre component, on the ISMIR'04 training set a 1NN accuracy of $\mathbf{88.8}\%$, and on the full ISMIR'04 set an accuracy of $\mathbf{91.8}\%$ is reached. On the HOMBURG set, 11NN classification using only the normalised timbre component yields $\mathbf{58.4}\%$.

Common sense indicates that the "unified" algorithm is a better choice for similarity estimation than such tuned variants, as the tuned variants do not perform well on all collections. In particular, these experiments show that discarding the rhythm component and using the timbre component alone, higher accuracies than those of the "unified" algorithm are obtained both on the ISMIR'04 set and the HOMBURG set. But with this setting, accuracy decreases clearly on the "Ballroom Dancers" collection. This may indicate the existence of an *evaluation glass ceiling* in the sense that an improved general music similarity algorithm might even yield lower accuracies.

5 CONCLUSIONS

We have presented modifications of Fluctuation Patterns (FPs) that can be used to obtain higher classification accuracies on the audio signal of the "Ballroom Dancers collection" than FPs compared by Euclidean distance. By adding frequency information to these proposed rhythm descriptors in the form of a "timbral" component results are further improved.

Based on these results, we suggest a "unified" algorithm. The presented experiments indicate that the similarities computed by this algorithm *both* reflect aspects of rhythm similarity *and* aspects of general music similarity. In both respects, classification accuracies obtained in our test setting are at least comparable to those previously reported for algorithms specifically designed for the respective tasks.

Going beyond this, presented preliminary results show

that by using different parameter settings (including selection of used OCs, and relative weighting of timbral and rhythm component) for different collections, the accuracies obtained with the "unified" algorithm can be further improved. As by doing so, one loses the generality of the algorithm, we refrain from further optimizations in this direction.

6 ACKNOWLEDGMENTS

This work is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung under project number L511-N15.

7 REFERENCES

[1] Jean-Julien Aucouturier and Francois Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[2] Simon Dixon, Fabien Gouyon, and Gerhard Widmer. Towards characterisation of music via rhythmic patterns. In *Proc. International Conference on Music Information Retrieval (ISMIR'04)*, 2004.

[3] A. Flexer, F. Gouyon, S. Dixon, and G. Widmer. Probabilistic combination of features for music classification. In *Proc. International Conference on Music Information Retrieval (ISMIR'06)*, 2006.

[4] Fabien Gouyon and Simon Dixon. Dance Music Classification: A Tempo-Based Approach. In *Proc. International Conference on Music Information Retrieval (ISMIR'04)*, 2004.

[5] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proc. AES* 25^{th} *International Conference*, 2004.

[6] A. Holzapfel and Y. Stylianou. Musical genre classification using nonnegative matrix factorization-based features. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):424–434, 2008.

[7] Andre Holzapfel and Yannis Stylianou. A scale transform based method for rhythmic similarity of music. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[8] Helge Homburg, Ingo Mierswa, Bülent Möller, Katharina Morik, and Michael Wurst. A benchmark dataset for audio classification and clustering. In *Proc. International Conference on Music Information Retrieval (ISMIR'05)*, 2005.

[9] Dan-Ning Jiang Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2002.

[10] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. International Conference on Music Information Retrieval (ISMIR'05)*, 2005.

[11] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151, 1991.

[12] Fabian Moerchen, Ingo Mierswa, and Alfred Ultsch. Understandable models of music collections based on exhaustive feature generation with temporal statistics. In *Proc.*12^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[13] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proc. International Conference on Music Information Retrieval (ISMIR'08)*, 2008.

[14] E. Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrieval*. Doctoral dissertation, Vienna University of Technology, 2006.

[15] Geoffroy Peeters. Rhythm classification using spectral rhythm patterns. In *Proc. International Conference on Music Information Retrieval (ISMIR'05)*, 2005.

[16] T. Pohle, P. Knees, M. Schedl, and G. Widmer. Automatically adapting the structure of audio similarity spaces. In *Proc.* 1^{st} *Workshop on Learning the Semantics of Audio Signals (LSAS 2006),* 1^{st} *International Conference on Semantics and Digital Media Technology (SAMT 2006)*, 2006.

[17] Yuan-Yuan Shi, Xuan Zhu, Hyoung-Gook Kim, Ki-Wan Eom, and Kim Ji-Yeun. Log-scale modulation frequency coefficient: A tempo feature for music emotion classification. In *Proc.* 1^{st} *Workshop on Learning the Semantics of Audio Signals (LSAS 2006),* 1^{st} *International Conference on Semantics and Digital Media Technology (SAMT 2006)*, 2006.

[18] Kris West. *Novel techniques for Audio Music Classification and Search*. Doctoral dissertation, School of Computing Sciences, University of East Anglia, 2008.

[19] Linxing Xiao, Aibo Tian, Wen Li, and Jie Zhou. Using a Statistic Model to Capture the Association Between Timbre and Perceived Tempo. In *Proc. International Conference on Music Information Retrieval (ISMIR'08)*, 2008.

[20] Wei Xu, Jacques Duchateau, Kris Demuynck, and Ioannis Dologlou. A New Approach to Merging Gaussian Densities in Large Vocabulary Continuous Speech Recognition. In *Proc. IEEE Benelux Signal Processing Symposium*, 1998.