EVALUATION OF FEATURE EXTRACTORS AND PSYCHO-ACOUSTIC TRANSFORMATIONS FOR MUSIC GENRE CLASSIFICATION

Thomas Lidy **Andreas Rauber**
Vienna University of Technology
Department of Software Technology and Interactive Systems
Favoritenstrasse 9-11/188, A-1040 Vienna, Austria
{lidy, rauber}@ifs.tuwien.ac.at

ABSTRACT

We present a study on the importance of psycho-acoustic transformations for effective audio feature calculation. From the results, both crucial and problematic parts of the algorithm for Rhythm Patterns feature extraction are identified. We furthermore introduce two new feature representations in this context: Statistical Spectrum Descriptors and Rhythm Histogram features. Evaluation on both the individual and combined feature sets is accomplished through a music genre classification task, involving 3 reference audio collections. Results are compared to published measures on the same data sets. Experiments confirmed that in all settings the inclusion of psycho-acoustic transformations provides significant improvement of classification accuracy.

Keywords: content-based retrieval, psycho-acoustic, audio feature extraction, music genre classification

1 INTRODUCTION

Digital music databases are continuously gaining popularity both in terms of professional repositories and personal audio collections. Ongoing advances in network bandwidth and popularity of internet services anticipate even further growth of the number of people involved with audio libraries. However, organization of large music repositories is a tedious and time-intensive task, especially when the traditional solution of manually annotating semantic data to the audio is chosen. Fortunately, research in music information retrieval has made substantial progress in recent years. Approaches from music information retrieval accomplish content-based audio analysis and are fundamental to tasks like browsing by similarity, automatic retrieval, organization or classification of music. Content-based descriptors form the base for these tasks and are able to add semantic meta-data to music. However, there is no absolute definition of what defines the content, or semantics, of a piece of audio. It is a matter of the specific application domain – and also of ongoing research – whether feature extractors lay their focus on musical elements such as timbre, pitch, tempo, energy distribution, rhythm or other content.

According to Aucouturier and Pachet (2003) musical genre is probably the most popular metadata for the description of music content. Music industry promotes the use of genres and home users like to organize their audio collections by this annotation. Consequently, the need of automatic classification of audio data into genres increased substantially, as did the number of researchers addressing this problem. Besides recent advances in genre classification there is still the question, what exactly defines a genre, or whether it is mainly dependent on a user's experience and taste. Aucouturier and Pachet (2003) deal with this question and the problem of inconsistent genre taxonomies.

Though the concept of musical genre might be ill-defined, recent approaches that use audio feature extraction combined with machine learning techniques achieve promising results. Genre classifiers typically work well with clearly described, well-distinguishable genres.

One of our main contributions to research in this area has been a feature extractor that describes rhythmic structure on a variety of frequency bands considering psycho-acoustic phenomenons according to human perception. The feature set, called Rhythm Patterns (RP), is neither a mere description of rhythm nor does it represent plain pitch information. Rather, it describes the modulation of the sensation of loudness for different bands, by means of a time-invariant frequency representation. We created an approach making the Rhythm Patterns feature set audible, enabling humans to get a notion of the calculated features (Lidy et al., 2005). An overview of the entire SOMeJB system is given by Neumayer et al. (2005).

One of the primary characteristics of the feature set is the integration of a range of psycho-acoustic processing steps. A question that was raised several times by reviewers and fellow researchers in this field was on the necessity and impact of these transformations. The replication of the human auditory system for computing similarity between signals was questioned. In this paper we address this issue, performing a range of experiments on 3 standard music IR reference collections, evaluating the

impact of the different psycho-acoustic processing steps. We furthermore introduce 2 new feature representations in this context and evaluate their performance both individually as well as in combination with the Rhythm Patterns features.

2 RELATED WORK

The domain of content-based music retrieval experienced a major boost in the late 1990's when mature techniques for the description of audio content became available. From that time on a range of researchers has been working on different methods for content-based retrieval. As manifold as the feature calculation approaches are the similarity measures and the evaluation methods. Here, we briefly review the major contributions on content-based feature extraction from audio.

One of the first works on content-based retrieval of audio (Foote, 1997) presents a search engine which retrieves audio from a database by similarity to a query sound. For similarity, two different distance measures are described in the paper.

An early work on musical style recognition (Dannenberg et al., 1997) investigates various machine learning techniques applied for building style classifiers.

Liu and Huang (2000) propose a new approach for content-based audio indexing using Gaussian Mixture Models and describe a new metric for distance measuring between two models. Logan and Salomon (2001) perform content-based audio retrieval based on K-Means clustering of MFCC features and define another novel distance measure for comparison of descriptors. Aucouturier and Pachet (2002) introduce a timbral similarity measure based on Gaussian Mixture Models of MFCCs, but also question the use of such measures in very large databases and propose a measure of "interestingness".

Pampalk et al. (2003) conduct a comparison of several content-based audio descriptors on both small and large audio databases, including those of Logan and Salomon (2001) and Aucouturier and Pachet (2002) as well as a feature set called Fluctuation Patterns, similar to the Rhythm Patterns we used in our experiments. They report that in the large scale evaluation the simple spectrum histograms outperform all other descriptors.

Li et al. (2003) propose Daubechies Wavelet Coefficient Histograms as a feature set suitable for music genre classification. The feature set characterizes amplitude variations in the audio signal. Experiments with several learning classifiers, including Support Vector Machines, have been conducted.

A large-scale evaluation with both subjective and content-based similarity measures was performed by Berenzweig et al. (2003). They addressed the question of comparing different existing music similarity measures and also raised the demand for a common evaluation database.

Basili et al. (2004) present a study on different machine learning algorithms (and varying dataset partitioning) and their performance in music genre classification.

Dixon et al. (2004) conduct experiments with parallels to ours: they utilize rhythmic patterns combined with additional features derived from them and evaluate on the same database as one of the three we used.

Facing the number of different approaches and evaluation measures, the call for common evaluation among the MIR research groups has grown substantially (Downie, 2003). Much effort has been put in organizing a Music IR contest, that was first held during ISMIR 2004, evaluating MIR performance in 5 different tasks, and which is now being continued as the MIREX contest (MIREX 2005).

3 FEATURE SETS

3.1 Rhythm Patterns Features

The Rhythm Patterns form the core of the SOM-enhanced JukeBox (SOMeJB) system, which was first introduced by Rauber and Frühwirth (2001) without any psycho-acoustic processing. The approach was later drastically enhanced by incorporating psycho-acoustic phenomena (Rauber et al., 2002). In the current incarnation of the feature set, audio at 44 kHz sampling resolution is processed directly, in mono format. Several improvements and code optimizations regarding processing time have been made and numerous options have been introduced, such as automatic choice of window step width. A number of the following steps which are carried out during audio feature extraction are now optional. The algorithm for extracting the Rhythm Patterns is as follows:

preprocessing 1 convert audio from au, wav or mp3 format to raw digital audio

preprocessing 2 if audio contains multiple channels, average them to 1 channel

preprocessing 3 take a 6 second excerpt from the audio, according to current processing position and considering lead-in, fade-out and step-width options

step [S1] transform audio segment into spectrogram representation using Fast Fourier Transform (FFT) with hanning window function (23 ms windows) and 50 % overlap

step [S2] apply Bark scale (Zwicker and Fastl, 1999) by grouping frequency bands into 24 critical bands

step [S3] apply spreading function to account for spectral masking effects (Schröder et al., 1979)

step [S4] transform spectrum energy values on the critical bands into decibel scale [dB]

step [S5] calculate loudness levels through incorporating equal-loudness contours [Phon]

step [S6] compute specific loudness sensation per critical band [Sone]

step [R1] apply Fast Fourier Transform (FFT) to the Sone representation. The result is a time-invariant representation of the 24 critical bands that captures reocurring patterns in the audio signal and thus is able to show rhythmic structure on each of the critical bands, i.e. amplitude modulation with respect to modulation frequencies. The transformation obtains amplitude modulation in the range from 0 to 43 Hz, however only the range from 0 through 10 Hz is considered in the Rhythm Patterns, as higher values are beyond what humans can perceive as rhythm.

step [R2] weight modulation amplitudes according to fluctuation strength sensation. According to human hearing sensation amplitude modulations are perceived most intense at 4 Hz and decreasing towards 15 Hz.

step [R3] apply a gradient filter to emphasize distinctive beats and perform Gaussian smoothing to increase similarity between two feature descriptors by diminishing un-noticeable variations.

postprocessing from all the Rhythm Patterns descriptors retrieved from the 6 second segments of a given piece of music, calculate the median as a descriptor for the whole piece of music

The steps [S2] through [S6] as well as [R2] incorporate psycho-acoustic phenomenons, based on studies of the human hearing system. Steps [S3], [S4], [S5], [S6], [R2] and [R3] can be performed optionally. It is their contribution to similarity representation that is of interest in this paper.

3.2 Statistical Spectrum Descriptor

During feature extraction we compute a Statistical Spectrum Descriptor (SSD) for the 24 critical bands. The spectrum transformed into Bark scale in step [S2] in Section 3.1 represents rhythmic characteristics within the specific frequency range of a critical band. According to the occurrence of beats or other rhythmic variation of energy on a specific band, statistical measures are able to describe the audio content. We intend to describe the rhythmic content of a piece of audio by computing the following statistical moments on the values of each of the 24 critical bands: mean, median, variance, skewness, kurtosis, min- and max-value. They can be calculated after any of the steps during Rhythm Patterns feature calculation, however we usually retrieve them after step [S2] or [S6]. The resulting Statistical Spectrum Descriptor contains 168 feature attributes.

3.3 Rhythm Histogram Features

The Rhythm Histogram features we use are a descriptor for general rhythmics in an audio document. Contrary to the Rhythm Patterns and the Statistical Spectrum Descriptor, information is not stored per critical band. Rather, the magnitudes of each modulation frequency bin of all 24 critical bands are summed up, to form a histogram of "rhythmic energy" per modulation frequency. The histogram contains 60 bins which reflect modulation frequency between 0 and 10 Hz. For a given piece of audio, the Rhythm Histogram feature set is calculated by taking the median of the histograms of every 6 second segment processed, resulting in a 60-dimensional feature space.

4 EXPERIMENTS

4.1 Audio collections and Experiment setup

We present a range of experiments performed on the Rhythm Patterns Feature Set, the Statistical Spectrum Descriptor and the Rhythm Histogram Features, as well as combinations of them. For a quantitive evaluation of each of the feature sets we measure their performance in classification tasks. The task is to classify the music documents into a predetermined list of classes, i.e. genres, according to a previously annotated ground-truth. The experiments were performed on three different audio collections in order to gain information about the generalization of the results to different music repositories and thus different musical styles, or to possibly detect specific problems with certain types of audio. The first audio collection is the one that was used by George Tzanetakis in previous experiments (Tzanetakis, 2002), consecutively denoted as GTZAN. It consists of 1000 pieces of audio equidistributed among 10 popular music genres. The second collection is the one used in the ISMIR 2004 Rhythm classification contest (ISMIR2004contest), which consists of 698 excerpts of 8 genres from ballroom dance music. The third collection is from the ISMIR 2004 Genre classification contest (ISMIR2004contest) and contains 1458 complete songs, the pieces being unequally distributed over 6 genres. For details about the genres involved in each collection and the numbers of documents in each class refer to Table 1.

Table 1: Three audio collections used in the experiments listing classes and number of titles per class.

GTZAN	1000	ISMIRrhythm	698	ISMIRgenre	1458
blues	100	ChaChaCha	111	classical	640
classical	100	Jive	60	electronic	229
country	100	Quickstep	82	jazz_blues	52
disco	100	Rumba	98	metal_punk	90
hiphop	100	Samba	86	rock_pop	203
jazz	100	SlowWaltz	110	world	244
metal	100	Tango	86		
pop	100	VienneseWaltz	65		
reggae	100				
rock	100				

For classification, we used Support Vector Machines with pairwise classification. A 10-fold cross validation was performed in each experiment from which we report macro-averaged precision and recall, defined as:

$$P^M = \frac{\sum_{i=1}^{|C|} \pi_i}{|C|}, \quad R^M = \frac{\sum_{i=1}^{|C|} \rho_i}{|C|} \tag{1}$$

where $|C|$ is the number of classes in a collection, and precision π_i and recall ρ_i per class are defined as:

$$\pi_i = \frac{TP_i}{TP_i + FP_i}, \quad \rho_i = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

where TP_i is the number of true positives in class i, FP_i is the number of false positives in class i, i.e. documents identified as class i but actually belonging to another class, and FN_i is the number of false negatives of a class i, i.e. documents belonging to class i, but which the classifier assigned to another class. We report macro-averaged precision and recall in order to make up for the unequal distribution of classes in the ISMIRgenre and ISMIRrhythm data collections. As globally comparable criterion we report the F_1 measure

$$F_1 = \frac{2 \cdot P^M \cdot R^M}{P^M + R^M} \tag{3}$$

which is a combined measure of precision and recall, attributing the same weight to both as it is their harmonic mean. Additionally, for comparability to other studies, we report Accuracy, defined as

$$A = \frac{\sum_{i=1}^{|C|} TP_i}{N} \quad (4)$$

N being the total number of audio documents in a collection.

4.2 Rhythm Patterns Variants

In the first series of our experiments we compared variations of our original algorithm for the extraction of the Rhythm Patterns features. Our specific interest is the impact of the various psycho-acoustic transformations. With the results from this experiments, we obtain information about the important parts of the feature extraction algorithm as well as an indication of which parts potentially pose problems to the performance of the feature set.

Table 2 provides an overview of the experiments. Each experiment is identified by a letter. The table lists the steps of the feature extraction process involved in each experiment. Experiment A represents the baseline, where all the feature extraction steps are involved. Experiments K through N completely omit the transformations into the dB, Phon and Sone scales. Experiments G to I and K to Q extract features from the audio without accounting for spectral masking effects. A number of experiments evaluates the effect of filtering/smoothing and/or the fluctuation strength weighting.

In Table 3 our results from experiments A through Q on the three audio collections are presented (best and second-best result in each column printed in boldface). From the results of the experiments we make several interesting observations. Probably the most salient observation is the low performance of the experiments J through N (with the exception of the precision in the ISMIRgenre collection). These experiments do not involve transformation into decibel scale nor successive transformation into the Phon and Sone scales. Also, experiments E and F as well as H and I deliver quite poor results, at least on the GTZAN and ISMIRgenre data sets. Those experiments perform decibel transformation but skip the transformation into Phon and/or Sone. All these results indicate clearly that transformation into the logarithmic decibel scale is very important, if not essential, for the audio feature extraction and subsequent classification or retrieval tasks. The successive application of the equal loudness curves (i.e. Phon transformation) and the calculation of Sone values appear also as important steps during feature extraction (experiment A compared to E and F, or experiment G compared to H and I).

Spectral Masking (i.e. step S3) was the subject of numerous experiments. We wanted to measure the influence of the use or omission of the spreading function for spectral masking together with variations in the other feature extraction steps. Table 3 clearly shows, that most experiments without Spectral Masking achieved better results. The ISMIRrhythm collection constitutes an exception to this. Nevertheless, the degradation of results incorporating spectral masking raises the question whether the spectral masking spreading function is inappropriate for music of certain styles.

Further focus of investigation were the effects of the fluctuation strength weighting curve (step R2) and the filtering/smoothing of the Ryhthm patterns (step R3). Both the GTZAN and ISMIRgenre collections perform significantly better with gradient filter and smoothing turned off. The ISMIRrhythm collection, however, shows contrary results. Its results improve when omitting the fluctuation strength weighting, but degrade when filtering & smoothing is omitted.

As we see in several experiments, the ISMIRrhythm collection behaves quite contrary to the two other collections. At this point we must note, that the overall results of the ISMIRrhythm collection are by far better than the ones carried out with the two other collections. The reason why this collection behaves differently might be that the results are already at a high level and variations in the algorithm only cause small fluctuations in the result values. On the other hand, contrary to the GTZAN collection and ISMIRgenre collection, ISMIRrhythm contains music from 8 different dances. The discrimination of ballroom dances relies heavily, if not exclusively, on rhythmic structure, which makes our Rhythm Patterns feature set an ideal descriptor (and thus justifies the good results). Apparently, smoothing the Rhythm Patterns is important for making dances from the same class with slightly different rhythms more similar – whereas in the two other collections, filtering & smoothing has negative effects. The ISMIRrhythm set appears to be independent of the spectral masking effects. Best results with ISMIRrhythm were retrieved with experiment C, which omits fluctuation strength weighting [R2], closely followed by experiment P, which additionally omits spectral masking [S3].

For the GTZAN and ISMIRgenre collections best results both in terms of F_1 measure and Accuracy were achieved in experiment O, which is the original Rhythm Patterns feature extraction without spectral masking [S3] and without filtering & smoothing [R3].

4.3 Statistical Spectrum Descriptor Experiments

In the experiments with the Statistical Spectrum Descriptor (SSD) we mainly investigate the performance of the features depending on which position in the Rhythm Patterns feature extraction process they are computed. Two positions were chosen to test the SSD: First, the statistical measures are derived directly after step [S2], when the frequency bands of the audio spectrogram have been grouped to critical bands. In the second experiment, the features are calculated after the critical bands spectrum had undergone logarithmic dB transformation as well as transformation into Phon and Sone, i.e. after step [S6]. In order to find an adequate representation of an audio track through a Statistical Spectrum Descriptor, we evaluated both the calculation of the mean and the median of all segments of a track.

Table 4 gives the results of the 4 experiment variants. From the results we find, that in any case the calculation after step [S6] is superior to deriving the SSD already at

Table 2: Experiment IDs and the steps of the Rhythm Patterns feature extraction process involved in each experiment.

step		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S1	FFT	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
S2	Critical bands	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
S3	Spectral masking	×	×	×	×	×	×				×							
S4	dB transform	×	×	×	×	×	×	×	×	×						×	×	×
S5	Equal loudness (Phon)	×	×	×	×	×		×	×							×	×	×
S6	Spec. loudness Sens. (Sone)	×	×	×	×			×								×	×	×
R1	FFT Modulation Amplitude	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
R2	Fluctuation Strength	×	×			×	×	×	×	×	×	×	×			×		
R3	filter/smoothing	×		×		×	×	×	×	×	×	×		×			×	

Table 3: Results of the Rhythm Patterns feature extraction experiments, for 3 audio collections, using 10-fold cross validation, in terms of macro-averaged precision (P^M), macro-averaged recall (R^M), F_1 measure and Accuracy (A). All values in %. Highest and second highest value in each column are boldfaced.

	GTZAN				ISMIRrhythm				ISMIRgenre			
Exp.	P^M	R^M	F_1	A	P^M	R^M	F_1	A	P^M	R^M	F_1	A
A	58.51	58.50	58.50	58.50	82.50	81.28	81.88	81.66	59.83	56.07	57.89	70.99
B	62.64	62.30	62.47	62.30	**83.35**	81.56	82.45	**82.38**	62.42	61.80	62.11	72.63
C	59.67	59.40	59.53	59.40	**83.39**	**82.30**	**82.84**	**82.81**	59.65	56.28	57.92	71.19
D	62.64	62.30	62.47	62.30	83.20	81.35	82.26	82.24	62.45	61.60	62.02	72.63
E	55.51	55.80	55.65	55.80	81.74	80.85	81.29	81.38	59.63	57.98	58.80	70.44
F	53.57	53.60	53.58	53.60	82.04	81.14	81.59	81.66	57.20	54.73	55.94	68.24
G	**62.96**	**62.90**	**62.93**	**62.90**	82.59	81.61	82.10	81.95	65.62	60.83	63.13	73.73
H	59.06	59.50	59.28	59.50	81.94	80.58	81.25	81.38	59.63	58.63	59.13	71.47
I	59.71	60.20	59.95	60.20	82.39	81.00	81.69	81.81	59.40	57.88	58.63	70.30
J	53.06	52.30	52.68	52.30	74.09	72.71	73.39	73.50	64.50	51.97	57.56	69.27
K	53.85	53.10	53.47	53.10	74.06	72.25	73.15	73.35	**66.80**	52.65	58.89	70.03
L	55.08	54.40	54.74	54.40	67.05	66.50	66.77	67.77	63.80	54.48	58.77	69.62
M	54.46	53.90	54.18	53.90	74.86	72.44	73.63	73.50	**66.40**	52.22	58.46	69.20
N	55.36	54.70	55.03	54.70	66.98	66.26	66.62	67.34	63.37	53.82	58.20	69.14
O	**64.22**	**64.40**	**64.31**	**64.40**	80.73	79.34	80.03	80.09	65.08	**64.50**	**64.79**	**75.03**
P	60.51	60.50	60.50	60.50	83.15	**81.88**	**82.51**	82.24	66.15	61.57	63.78	73.94
Q	**64.22**	**64.40**	**64.31**	**64.40**	81.58	80.16	80.86	80.95	64.87	**64.13**	**64.50**	**74.90**

the earlier stage [S2]. As in the experiments with the Rhythm Patterns feature set, logarithmic transformation appears to be essential for the results of the content-based audio descriptors. Comparing the summarization of an audio track by mean and by median, results of the GTZAN and ISMIRgenre collection argue for the use of the mean. Again, the ISMIRrhythm collection indicates contrary results, however the differences in result measures vary only between 0.04 and 1.4 percentage points.

Note, that the SSD feature set calculated after step [S6] outperforms the Rhythm Patterns descriptor both in the GTZAN and ISMIRgenre collections. This is especially remarkable as the statistical descriptors have a dimensionality 8.5 times lower than the Rhythm Patterns feature set.

4.4 Experiments on Rhythm Histogram Features

The Rhythm Histogram Features (RH) describe global rhythmic content of a piece of audio by a measure of energy per modulation frequency. They are calculated from the time-invariant representation of the Rhythm Patterns. Our experiments tried to evaluate different performance when computing the Rhythm Histogram Features after feature extraction step R1, R2 or R3, respectively. Evaluation showed, that regardless to the stage, RH features virtually always produce equal results. We thus omit a table with detailed results; performance of the Rhythm Histogram features can be seen in the row denoted 'RH [R1]' of Table 5.

Results of the RH features in the ISMIRrhythm collection achieve nearly the results of the Rhythm Patterns feature set. Note that dimensionality is 24 times lower than that of the latter one. Performance of GTZAN and ISMIRgenre collections is rather low, nevertheless, though being a simple descriptor, the Rhythm Histogram feature set seems eligible for audio content description.

4.5 Comparison and Combined Feature sets

Table 5 displays a comparison of the baseline Rhythm Patterns (RP) algorithm (experiment A) to the best results

Table 4: Results of the experiments with Statistical Spectrum Descriptor (3 data sets, 10-fold cross val., best results bold).

	GTZAN				ISMIRrhythm				ISMIRgenre			
Exp.	P^M	R^M	F_1	A	P^M	R^M	F_1	A	P^M	R^M	F_1	A
SSD[S2] (mean)	60.87	60.20	60.53	60.20	36.56	21.08	26.74	25.64	40.58	25.57	31.37	51.58
SSD[S2] (median)	57.70	57.00	57.35	57.00	43.54	39.96	41.67	43.84	68.17	49.90	57.62	67.76
SSD[S6] (mean)	**72.85**	**72.70**	**72.77**	**72.70**	54.35	52.81	53.57	54.73	**76.93**	**67.95**	**72.16**	**78.53**
SSD[S6] (median)	71.57	71.30	71.43	71.30	**54.39**	**53.80**	**54.09**	**55.44**	75.78	66.70	70.95	77.50

of the Rhythm Patterns extraction variants, the Statistical Spectrum Descriptor (SSD) and the Rhythm Histogram features (RH). Best results in Rhythm Patterns extraction were achieved with the GTZAN, ISMIRrhythm and ISMIRgenre audio collections in experiments O, C, and O respectively. Accuracy was 64.4, 82.8, and 75.0 %, respectively. The Statistical Spectrum Descriptor performed best when calculated after psycho-acoustic transformations, and taking the simple mean of the segments of a piece of audio. Accuracy was 72.7, 54.7, and 78.5 % in the GTZAN, ISMIRrhythm and ISMIRgenre data set, respectively, which exceeds the Rhythm Patterns feature set in 2 of the 3 collections. Rhythm Histogram Features achieved 44.1, 79.94, and 63.17 % accuracy, which rival the Rhythm Patterns features regarding the ISMIRrhythm data collection. Obviously a combination of feature sets offers itself for further improvement of classification performance.

Various experiments on 2 set combinations have been evaluated. The combination of Rhythm Patterns features with the Statistical Spectrum Descriptor achieves 72.3 % accuracy in the GTZAN data set, which is slightly lower than the performance of the SSD alone. Contrary, in the ISMIRrhythm data set, the combination achieves a slight improvement. In the ISMIRgenre audio collection, this combination results in a significant improvement and achieves the best result of all experiments on this data set (80.32 % accuracy).

Combination of Rhythm Patterns features with Rhythm Histogram Features changes the results of the Rhythm Patterns features only insignificantly, a noticeable improvement can be seen only in the ISMIRrhythm data set, which is the data set where the Rhythm Histogram features performed best.

Very interesting are the results of combining the Statistical Spectrum Descriptor with Rhythm Histogram features: With the GTZAN collection, this combination achieves the best accuracy (74.9 %) of all experiments (including the 3 set experiments). The result on the ISMIRrhythm collection is comparable to the best Rhythm Patterns result. The 2 set combination without Rhythm Patterns features performs also very well on the ISMIRgenre data set, achieving the best F_1 measure (73.3 %). There is a notably high precision value of 76.67 %, however, recall is only at 70.22 %. Accuracy is 79.63 % and thus slightly lower than in the Rhythm Patterns + SSD combination.

Finally, we investigated the combination of all 3 feature sets, which further improved the results only on the ISMIRrhythm data set. Accuracy increased to 84.24 %, compared to 82.81 % using only the Rhythm Patterns features. As stated, results on the ISMIRrhythm collection were rather high from the beginning, consequently improvements on classification in this data set were moderate.

Overall improvement, regarding best accuracy values achieved in each data collection compared to baseline experiment A, was +16.4 percentage points on the GTZAN music collection, +2.58 percentage points on the ISMIRrhythm collection and +9.33 percentage points on the ISMIRgenre music collection.

4.6 Comparison with other results

4.6.1 GTZAN data set

The GTZAN audio collection was assembled and used first in experiments by Tzanetakis (2002). The original collection was organized in a three level hierarchy intended for discrimination into speech/music, classification of music into 10 genres and subsequent classification of the two genres classical and jazz into subgenres. In our experiments we used the organization of 10 musical genres in the second level, and thus compare our results to the performance of Tzanetakis (2002) on that level. The best classification result reported was 61 % accuracy (4 % standard deviation on 100 iterations of a 10-fold cross validation) using Gaussian Mixture Models and the 30 dimensional MARSYAS genre features.

Li et al. (2003) used the same audio collection in their study and compare "Daubechies Wavelet Coefficient Histograms" (DWCHs) to combinations of MARSYAS features. DWCHs achieved 74.9 % classification accuracy in a 10-fold cross validation using Support Vector Machines (SVM) with pairwise classification and 78.5 % accuracy using SVM with one-versus-the-rest classification.

Our current best performance is 74.9 %, which constitutes an improvement of 16.4 percentage points regarding original Rhythm Patterns feature descriptor.

Table 6: Comparison with other results on the GTZAN audio collection (10-fold cross validation).

GTZAN	A
Tzanetakis (2002) (GMM)	61.0
Li et al. (2003) (SVM pairwise)	74.9
Li et al. (2003) (SVM one-vs-the-rest)	78.5
our best result (SVM pairwise)	74.9

4.6.2 ISMIRrhythm data set

Though not participating in the ISMIR Rhythm classification contest, two papers of ISMIR 2004 report experiment

Table 5: Comparison of feature sets and combinations (3 data sets, 10-fold cross validation, best results boldfaced).

	GTZAN				ISMIRrhythm				ISMIRgenre			
Exp.	P^M	R^M	F_1	A	P^M	R^M	F_1	A	P^M	R^M	F_1	A
RP(A)	58.51	58.50	58.50	58.50	82.50	81.28	81.88	81.66	59.83	56.07	57.89	70.99
RP(best) O/C/O	64.22	64.40	64.31	64.40	83.39	82.30	82.84	82.81	65.08	64.50	64.79	75.03
SSD [S6] (mean)	72.85	72.70	72.77	72.70	54.35	52.81	53.57	54.73	**76.93**	67.95	72.16	78.53
RH [R1]	43.55	44.10	43.82	44.10	82.09	79.14	80.59	79.94	41.58	39.20	40.36	63.17
RP(best)+SSD	72.17	72.30	72.23	72.30	84.38	82.88	83.62	83.52	72.33	**72.00**	72.17	**80.32**
RP(best)+RH	64.06	64.20	64.13	64.20	84.45	83.08	83.76	83.67	65.27	64.55	64.91	75.51
SSD+RH	**74.79**	**74.90**	**74.84**	**74.90**	83.13	81.44	82.27	82.66	76.67	70.22	**73.30**	79.63
RP(best)+SSD+RH	72.25	72.40	72.32	72.40	**85.00**	**83.43**	**84.21**	**84.24**	71.85	71.27	71.56	79.97

results on the same data collection. The approach used by Gouyon and Dixon (2004) is based on tempo probability functions for each of the 8 ballroom dances and successive pairwise or three-class classification and reports 67.6 % overall accuracy.

Dixon et al. (2004) specifically address the problem of dance music classification, and achieve an astounding result of 96 % accuracy when using a combination of various feature sets. Besides soundly elaborated descriptors, the approach also incorporates a-priori knowledge about tempo and thus drastically reduces the number of possible classes for a given audio instance.

The ground-truth-tempo approach has been previously described by Gouyon et al. (2004), where classification based solely on the pre-annotated tempo attribute reached 82.3 % accuracy (k-NN classifier, k=1). The paper also describes a variety of descriptor sets and reports 90.1 % accuracy on the combination of MFCC-like descriptors with ground-truth tempo and 78.9 % accuracy when using computed tempo instead.

All results presented in Table 7 have been evaluated through a 10-fold cross validation, except for the first one, which used the ISMIR contest training/test set split.

Table 7: Comparison with other results on the ISMIRrhythm audio collection (10-fold cross validation).

ISMIRrhythm	A
Lidy et al. in (ISMIR2004contest)	82.0
Gouyon and Dixon (2004)	67.6
Gouyon et al. (2004) wo/tempo-gt.	78.9
Gouyon et al. (2004) w/tempo-gt.	90.1
Dixon et al. (2004) wo/tempo-gt.	85.7
Dixon et al. (2004) w/tempo-gt.	96.0
our current best result	84.2

4.6.3 ISMIRgenre data set

The ISMIRgenre data set was assembled for the ISMIR 2004 Genre classification contest. Results from the Genre classification contest are shown in Table 8 in terms of Accuracy, and Accuracy normalized by the genre frequency (which is equal to macro-averaged Recall). In order to be able to compare our current results to the values stated in the table, instead of a 10-fold cross-validation we repeated our experiment with the combination of RP(O)+SSD features using the same training and test set partitioning as in the contest. Though not surpassing the winner of the 2004 contest, the results of our current evaluation represent a substantial improvement to the approach submitted to the 2004 contest, making it theoretically rank second place.

Table 8: Comparison with the results from the ISMIR 2004 Genre classification contest (50:50 training and test set split).

ISMIRgenre	A	A (norm.)
Thomas Lidy and Andreas Rauber	70.4	55.7
Dan Ellis and Brian Whitman	64.0	51.0
Kris West	78.3	67.2
Elias Pampalk	82.3	78.8
George Tzanetakis	71.3	58.6
our current approach	79.7	70.4

5 SUMMARY

We performed a study on the contribution of psychoacoustic transformations in the calculation of Rhythm Patterns for efficient content-based music description. Numerous experiments have been arranged to identify the important parts in the feature extraction process. Moreover, two additional descriptors calculated together with the Rhythm Patterns – namely the Rhythm Histogram features and the Statistical Spectrum Descriptor – were presented, and evaluated in their efficiency compared to other feature sets. Performance on all experiments was measured by the results in a music genre classification task. The feature sets, besides being suitable for music similarity retrieval, are intended to perform automatic organization tasks by classification into different semantical genres. In order to be able to assess the general applicability in various genre taxonomies, three different standard MIR audio collections have been used in the evaluation. Besides measuring the performance of each individual feature set, we investigated whether combinations of the feature sets would significantly increase the results. Compared to the original Rhythm Patterns audio descriptor, the experiments on the three music collections achieved accuracy improvements of 16.4, 9.33, and 2.58 percentage points, respectively.

Evaluation of the Rhythm Patterns experiment variants showed that the implementation of spectral masking in the feature extraction might pose a potential issue in

the audio description, at least regarding specific types of music. Furthermore, filtering and smoothing procedures as well as the weighting of fluctuation strength have been identified to have quite unpredictable influence in audio classification for different taxonomies. However, a series of psycho-acoustic transformations, namely the transformation into the logarithmic dB scale, equal loudness in the Phon scale and specific loudness sensation in terms of the Sone scale, has been identified to be crucial for the audio description task.

Future tasks involve further investigation of the filtering and weighting processes as well as their influence depending on varying audio repositories.

References

J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, October 2002.

J.-J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

R. Basili, A. Serafini, and A. Stellato. Classification of musical genre: a machine learning approach. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, October 2003.

R. B. Dannenberg, B. Thom, and D. Watson. A machine learning approach to musical style recognition. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 344–347, Thessaloniki, Greece, September 25-30 1997.

S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 509–516, Barcelona, Spain, October 2004.

J. S. Downie. Toward the scientific evaluation of music information retrieval systems. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Baltimore, Maryland, USA, October 26-30 2003.

J. T. Foote. Content-based retrieval of music and audio. In *Proceedings of SPIE Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–147, 1997.

F. Gouyon and S. Dixon. Dance music classification: A tempo-based approach. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, pages 196–204, London, UK, June 17-19 2004.

ISMIR2004contest. ISMIR 2004 Audio Description Contest. Website, 2004. `http://ismir2004.ismir.net/ISMIR_Contest.html`.

T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282 – 289, Toronto, Canada, 2003.

T. Lidy, G. Pölzlbauer, and A. Rauber. Sound re-synthesis from rhythm pattern features - audible insight into a music feature extraction process. In *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Spain, September 5-9 2005.

Z. Liu and Q. Huang. Content-based indexing and retrieval-by-example in audio. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, New York, USA, July 30 - Aug. 2 2000.

B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Tokyo, Japan, August 2001.

MIREX 2005. 2nd annual Music Information Retrieval Evaluation eXchange. Website, 2005. `http://www.music-ir.org/mirexwiki/index.php/Main_Page`.

R. Neumayer, T. Lidy, and A. Rauber. Content-based organization of digital audio collections. In *Proceedings of the 5th Open Workshop of MUSICNETWORK*, Vienna, Austria, July 4-5 2005.

E. Pampalk, S. Dixon, and G. Widmer. On the evaluation of perceptual similarity measures for music. In *Proceedings of the International Conference on Digital Audio Effects (DAFx-03)*, pages 7–12, London, UK, September 8-11 2003.

A. Rauber and M. Frühwirth. Automatically analyzing and organizing music archives. In *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Darmstadt, Germany, September 4-8 2001.

A. Rauber, E. Pampalk, and D. Merkl. Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles. In *Proceedings of the International Conference on Music Information Retrieval*, pages 71–80, Paris, France, October 13-17 2002.

M. Schröder, B. Atal, and J. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66:1647–1652, 1979.

G. Tzanetakis. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. PhD thesis, Computer Science Department, Princeton University, 2002.

E. Zwicker and H. Fastl. *Psychoacoustics - Facts and Models*, volume 22 of *Springer Series of Information Sciences*. Springer, Berlin, 1999.