BEAT HISTOGRAM FEATURES FOR RHYTHM-BASED MUSICAL GENRE CLASSIFICATION USING MULTIPLE NOVELTY FUNCTIONS

Athanasios Lykartsis
Audio Communication Group
Technische Universität Berlin
Berlin, Germany
athanasios.lykartsis@tu-berlin.de

Alexander Lerch
Center for Music Technology
Georgia Institute of Technology
Atlanta, Georgia, US
alexander.lerch@gatech.edu

ABSTRACT

In this paper we present beat histogram features for multiple level rhythm description and evaluate them in a musical genre classification task. Audio features pertaining to various musical content categories and their related novelty functions are extracted as a basis for the creation of beat histograms. The proposed features capture not only amplitude, but also tonal and general spectral changes in the signal, aiming to represent as much rhythmic information as possible. The most and least informative features are identified through feature selection methods and are then tested using Support Vector Machines on five genre datasets concerning classification accuracy against a baseline feature set. Results show that the presented features provide comparable classification accuracy with respect to other genre classification approaches using periodicity histograms and display a performance close to that of much more elaborate up-to-date approaches for rhythm description. The use of bar boundary annotations for the texture frames has provided an improvement for the dance-oriented Ballroom dataset. The comparably small number of descriptors and the possibility of evaluating the influence of specific signal components to the general rhythmic content encourage the further use of the method in rhythm description tasks.

1. INTRODUCTION

The extraction of features describing musical rhythm is an important and a challenging topic in Music Information Retrieval (MIR) with applications in many areas [1, 2, 3, 4]. Rhythm features can be of importance, for example, in musical genre classification, where they have been shown to improve classification accuracy and, in specific cases (such as for special, dance-music oriented datasets) allow very successful rhythm-based classification or similarity computation [2, 5, 6, 7, 4]. Since the concept of 'genre' can be related to various musical parameters [8, 9, 5], it is generally attempted to include features from all possibly relevant musical categories (timbre, pitch, loudness, and rhythm) in order to achieve good classification results. In this paper, we chose to focus on rhythm features for musical genre classification since computational rhythm description remains a challenging subject, even when considerable advances have been made lately in this field [7, 10, 11, 4]. A common approach to the description of the rhythmic content of a musical track focuses on the extraction of features based on changes in the signal's amplitude envelope, possibly applying frequency-band filtering or attempting to track energy changes from the signal's short time spectrum. The basic assumption behind this approach is that rhythm is strongly related to long-term amplitude periodicities and their statistical properties [2, 5, 6]. This assumption is, in our view, well-founded, since many music theoretical works on rhythm have stressed the multidimensional nature of rhythm, establishing the need to consider different musical properties and their temporal evolution when attempting to represent the rhythmic content of music [12, 13, 14, 15].

Several approaches for the automatic extraction of features describing rhythmic content for genre classification through a periodicity distribution representation have been proposed [16, 17, 18]. They are based on earlier studies on beat analysis [19], which aimed at creating a very low frequency periodicity representation (later commonly dubbed the *beat histogram* [16] or *periodicity histogram* [18]), used to extract musical parameters such as tempo as well as low-level features (e.g., statistical properties of the histogram such as mean or standard deviation). Traditionally, a measure of the signal amplitude envelope or its change over time is utilized as the novelty function for the extraction of a beat histogram [18, 16, 17]. Genre classification systems based on such representations have generally shown promising results, although the rhythm features did not perform as well as other subsets of the feature set such as timbre features [5, 16, 17]. There still exist only relatively few studies [20, 21, 18, 22, 3] where rhythm features alone were so efficient as to achieve high accuracy in a genre classification or similarity task. The dataset used in most of those cases (Ballroom [23]) includes only tracks from dance genres with presumably clearly distinguishable rhythmic properties, leading to very high performance but providing little information as to the suitability of the features for music with less distinctive rhythm.

Additionally, there exist many approaches for rhythm description which do not only base themselves on direct feature extraction from a periodicity representation, but rely on direct similarity measurements, computing distances between rhythm representations also being based on periodicities present in the signal [24, 25, 26, 7]. Finally, latest works have attempted to model rhythm using probabilistic models [10, 3, 27] or deep neural networks [11, 28]. Such methods have shown excellent results in rhythmic classification and similarity computation tasks (with accuracies ranging up to 95% [3, 27, 11]), attesting to the plausibility of addressing rhythm-based genre classification tasks with rhythmic similarity methods or elaborate rhythm modeling [4]. However, we identify two main drawbacks with this category of approaches: first, the interpretability of the features describing rhythm is limited, since the method of their generation is either too complex or based on purely technical and not music-theoretical considerations. This is not a problem, of course, if the goal is to develop a method providing high accuracy; it does, however, limit the possible conclusions which can be drawn from the classification task in itself and the features involved in it. Secondly, the complexity of the methods makes them prone to errors and random influences. Furthermore, the very high results achieved by some studies, although proving the suitability of such classifier systems for e.g. commercial use, should in our view be seen in a critical light for two reasons: on the one hand, it is questionable whether systems based only on rhythmic features could theoretically achieve such performance, since listening experiments show that even human subjects cannot distinguish genres perfectly [29]; on the other hand, it has been demonstrated very often in the intrinsically fuzzily defined genre classification task that complex systems relying on elaborate transformations and trained with small amounts of data might provide very high results, but in effect do not generalize to real-world data (a case of *overfitting*) or do not describe the quantities they are supposed to, their performance being an artifact of an erroneous ground truth or highly dataset-specific features [30, 31, 32]. We chose to focus on the periodicity representation methods and the features which can be extracted on their basis, conducting a detailed examination of the features which can be used in their context and their behavior for many different datasets. This approach allows to investigate the merits of those methods in depth and to identify which signal-based features bear the most importance for rhythmic description.

A common element of previous studies using periodicity representations allowing rhythmic content feature extraction for genre classification is that the features are derived from a beat histogram created on the basis of the signal amplitude and energy changes, as for example in [16, 17, 18]. This approach might seem intuitive at first, since it is based on the assumption that rhythmic properties are conveyed through amplitude or energy changes in the musical signal, which is a common consequence of rhythm perception modalities [33, 34]. However, relevant literature in music theory [12, 15, 14] and cognition [35, 36] indicates that rhythm arises as the combination or interplay of periodicities associated with different sound properties such as *amplitude* (e.g., accents), *spectral* (e.g., instrumentation changes) and *tonal* changes (e.g., chord changes). In the field of onset detection, novelty functions have been proposed which take into account spectral content changes [37, 38, 39]. The goal of this work is to capitalize on this observation by extracting novelty functions from different signal properties (which will allow to take not only amplitude-based but also spectral and tonal changes and their related periodicities into account), using the beat histogram method to create a basis for features describing the rhythmic content of the signal. We investigate the impact of using various musical qualities as novelty functions for beat histogram calculation on classification accuracy, identify the most and least descriptive features and feature groups, and compare the results to those of a traditional timbre-related feature set. Furthermore, the effect of bar boundary annotations from manually annotated data is investigated, since beat histograms based on musically more meaningful data are expected to produce more qualitative features. Finally, we examine the misclassifications in the confusion matrices to draw conclusions on the suitability of the features for rhythm-based classification and possible shortcomings.

The paper is structured as follows. In Section 2, the feature extraction procedure is described. In the third Section, the evaluation of the proposed features is given, along with information about feature selection and the datasets. In the fourth Section, results are presented and discussed in Section 5. Finally, we give conclusions and suggestions for future work (Section 6).

2. METHOD

Novelty functions are generally defined as temporal curves designating prominent changes of a signal [40, 41, 42, 43], resulting from a reduced or filtered version of the original signal from which the first difference is computed, in order to accentuate substantial changes in the monitored quantity. In this paper, we expand this definition somewhat and consider every temporal trajectory of a signal feature to effectively be a novelty function in order to use it as a basis for the beat histogram calculation. This assumption is justified from a practical point of view, since prominent changes in the magnitude of a signal feature (such as, e.g., spectral flux) are still represented (but not as accentuated), their inherent periodicities detectable by methods such as a Discrete Fourier Transform (DFT), an Auto-Correlation Function (ACF) or a resonant filter bank. Furthermore, the avoidance of taking the first difference function has the added advantage of reducing noise in the feature temporal trajectory. For rhythm analysis, signal characteristics beyond amplitude are included here since their periodicities also contribute to the overall rhythm. Examples include changes in instrumentation, which cause a universal, broadband change in the spectral content of a signal, or chord changes, which can be tracked by changes in the instantaneous pitch content.

Fig. 1 shows the novelty function (top) and the resulting beat histogram (bottom) for two features representing spectral change (Spectral Flux) and spectral shape (Spectral Flatness), respectively. The audio track is an excerpt from the *disco* class of the GTZAN dataset [16] and the features are extracted over the whole length of a single texture frame (3 s). In the excerpt, the drum section plays a straight $4/4$ measure with a beat each 0.48 s, clearly visible in the spectral flux temporal curve (tracking general spectral change) and as the prevalent periodicity (126 BPM) in the corresponding beat histogram. The changes tracked by the spectral flatness measure, however, more strongly reflect tonalness changes, taking place every 0.96 s, with a main periodicity of 63 BPM. Both the general form (distribution) of the histogram and the strength and exact BPM value of the most prevalent periodicities can be seen to differ significantly in the two examples. Based on this example, it is obvious that the changes tracked by different novelty functions can lead to different beat histograms for the same audio excerpt, with each histogram potentially providing a rhythmically meaningful description.

The novelty functions used in this paper cover envelope, spectral shape, and tonal content changes. The selected features are the following (for details on their computation, see [44]):

- **Spectral Shape** features include Spectral Flux (SF), Spectral Centroid (SCD), Mel Frequency Cepstral Coefficients (MFCC 1-13) and Spectral Flatness (SFL).
- **Tonal** novelty features comprise Spectral Tonal Power Ratio (STPR) and the Pitch Chroma Coefficients (SPC 1-12).
- **Envelope** novelty is tracked through the Root Mean Square (RMS) measure of the signal amplitude.

The beat histogram computation is similar to the one proposed by Tzanetakis [16] without implementation of the wavelet filtering. All features are calculated through a Short-Time-Fourier-Transform (STFT), except for the RMS which is extracted with the same temporal resolution parameters from the time domain signal. The complete procedure for the generation of a feature vector representing each track includes the following steps:

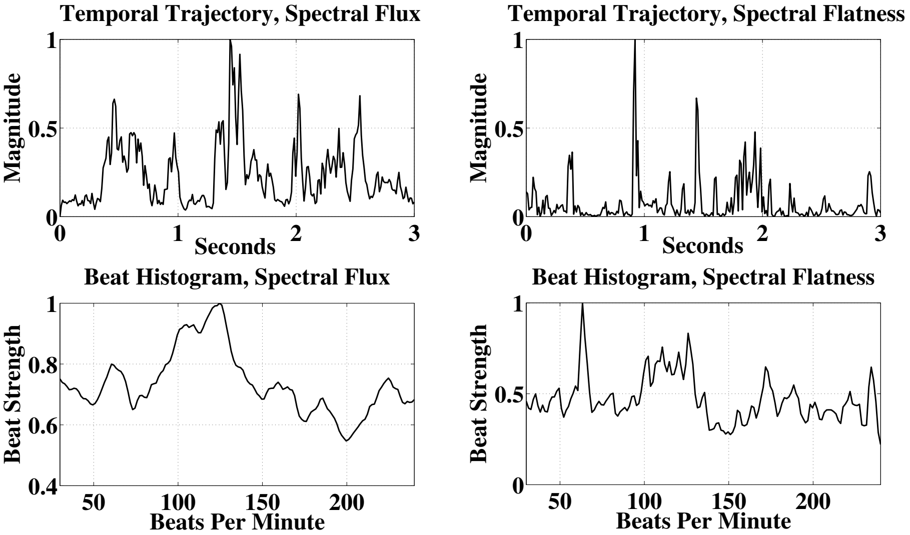

Figure 1: *Temporal Trajectory of Spectral Flux and Spectral Flatness (upper row) and corresponding Beat Histograms from* $30-240$ *BPM (lower row).*

1. **Preprocessing**: the audio signal is down-mixed to mono, resampled to 22.5 kHz, lowpass-filtered to remove DC components and normalized.

2. **STFT**: the transform is computed with a frame-length of approximately 46.4 ms, windowed by a Hann window and with 75% overlap between consecutive frames. The resulting frequency resolution is $\Delta f = 10.77$ Hz.

3. **Novelty function**: the features listed earlier are extracted from the STFT frames, and the novelty function is computed through the calculation of the temporal trajectory and half-wave rectification, as is commonly practiced in novelty function extraction [43].

4. **Beat histogram**: the beat histogram is extracted through the calculation of an Autocorrelation Function (ACF) for each texture window of length 3 s. The overlap of the texture windows is 75%.

5. **Subfeature computation**: for each beat histogram, the subfeatures listed in Table 1 are extracted. The concatenation of all subfeature groups for each novelty function produces the final feature vector for an audio excerpt.

Similar subfeatures can be found e.g. in [16] (Peak), and [17, 18] (Distribution). In total, 30 novelty functions are used for the production of as many beat histograms, from each of which 19 subfeatures are extracted, resulting in a total count of 570 features. This effectively means that from the temporal trajectory of every MFCC or chroma coefficient, as well as the other features mentioned in the previous paragraph, a beat histogram is extracted, ensuring that all relevant periodicities and their properties — in different frequency bands and describing various audio aspects — are accounted for.

Table 1: *Subfeatures extracted from Beat Histograms.*

Distribution	Peak
Mean (ME)	Salience of Strongest Peak (A1)
Standard Deviation (SD)	Salience of 2nd Stronger Peak (A0)
Mean of Derivative (MD)	Period of Strongest Peak (P1)
SD of Derivative (SDD)	Period of 2nd Stronger Peak (P2)
Skewness (SK)	Period of Peak Centroid (P3)
Kurtosis (KU)	Ratio of A0 to A1 (RA)
Entropy (EN)	Sum (SU)
Geometrical Mean (GM)	Sum of Power (SP)
Centroid (CD)	
Flatness (FL)	
High Frequency Content (HFC)	

3. EXPERIMENTAL SETUP AND EVALUATION

In order to be able to compare the results for the rhythm content features we extracted a *baseline feature set* by calculating the feature value over all texture windows of an excerpt (keeping the average value inside a window) without extracting periodicities. Those features are considered as a baseline since they represent a standard set of features used in genre classification. They include all features listed in Sect. 2 except the *Pitch Chroma Coefficients* and additionally include the features *Spectral Spread, Peak Amplitude Value*, and *Zero Crossing Rate*. The subfeatures on each of these features' temporal trajectory throughout each track are given in the *Distribution* column of Table 1. In total, the baseline feature set comprises 21 features times 11 subfeatures $=$ 231 features. We chose not to include other, more state-of-the-art non-rhythmic features in the baseline mainly because we aimed at investigating rhythmic features rather than non-rhythmic features. Second, since the main goal of the paper was to evaluate the beat histogram features derived from all relevant novelty functions of the signal, it seemed plausible to include as a baseline the same novelty functions. This, however, was performed without applying the beat histogram transformation, so as to assess the added value of this certain processing step and the resulting features.

3.1. Classification

For the classification part, we apply the Support Vector Machines (SVM) [45] algorithm under MATLAB with a Radial Basis Function (RBF) kernel. The two hyperparameters for this kernel C and γ were determined with a grid search procedure. For all the experiments presented here, a 10-fold cross-validation took place, and results are averaged over the folds. The goal of the given classification setup is to compare the performance of the rhythm content feature set to a standard set of features, while also testing the combined set in order to assess the improvement when using the rhythm feature set in addition to the baseline set. All features are subjected to standardization (z-score) prior to classification (train and test set separately). The performance measure used to evaluate the classification is *accuracy*, defined as the proportion of correctly classified samples to all samples classified.

3.2. Datasets

In order to ensure comparability of the results with other publications and to evaluate the rhythm features for different genre hierarchies and tracks, five datasets were evaluated:

- GTZAN [16]
- Ballroom [18, 23]
- ISMIR04 [46]
- Unique [47]
- Homburg [48]

Although none of the datasets raises claims to being either complete or error-free (since the definition of genre itself is a debatable subject), their previous extensive use makes them suitable as benchmarks for the musical genre classification task. In order to avoid having results which could be artifacts of one specific dataset (e.g., for the GTZAN which has been criticized for its content [31], or the Ballroom which has been seen to be easily classifiable by the tempo feature [18]), we chose to include five very diverse datasets here. An interesting task which came under consideration would be training with one dataset and testing with another. However, the different class/genre structure across them did not allow for such a use in this context.

3.3. Feature Selection

The large number of resulting features mentioned in Sect. 2 makes direct feature evaluation a tedious task. In order to identify the best and worst performing descriptors we conducted a two stage feature selection: First, we apply a filter method (Mutual Information with Target Data [49], using the maximum relevance CMIM metric [50] from the MI-Toolbox [51]). Second, we run a sequential forward feature selection using the SVM as a wrapper method [52]. We retain the N best features (the feature number N is dataset-dependent, but ranges between 10 and 20) which gave comparable to or better accuracy than the full feature sets. This procedure is applied to the baseline and rhythm feature sets separately, and the best features from both are then pooled to produce the combined feature set. Finally, we apply feature selection by separating feature subsets in the rhythm set in order to determine the effectiveness of different novelty functions and subfeature groups.

3.4. Bar Annotation

The parameters for the rhythm feature extraction algorithm were given in Sect. 2. It is a common problem of many such algorithms that the overlapping texture windows used in the block-wise processing of the audio file are of pre-defined length that does not necessarily represent "meaningful" parts of the music, such as, e.g., a bar. Placing the frames exactly at the boundaries of a bar or a musical phrase could increase the precision of the beat histogram representation, since the periodicities extracted from the segment would be musically meaningful, without onsets added or being left out because of random framing. In order to adapt the texture window boundaries to the bar boundaries, annotations of the audio files are necessary. In the case of the Ballroom dataset, such a manual annotation is available [53] and will be used here.

4. RESULTS

Results of the classification after feature selection for all datasets are presented in Fig. 2 and Table 2. The *priors* (percentage of a class/genre samples in a dataset) as average (Avg) and of the greatest class (Max P) are also provided. Sample results of the feature selection process concerning feature ranking (the three best and three worst features per dataset) are given in Table 3. Apart from that, in Fig. 3, accuracy results for all novelty function and subfeature groups for each dataset can be seen. Finally, confusion matrices are provided for the GTZAN and Ballroom datasets (Tables 4 and 5) for the rhythmic feature set in order to examine the misclassifications for those very well-known datasets.

Some tendencies can be clearly identified: The baseline feature set performs always better than the rhythm feature set alone except for the Ballroom dataset. The difference is mostly small but also significant at the 0.05 level in all cases except for the ISMIR04 dataset (based on a comparison test of the Cohen's Kappa extracted from the confusion matrices [54]); it ranges from 1.9% for the ISMIR04 dataset to 12.3% for the GTZAN dataset. Only for the Ballroom dataset the accuracy using the rhythm feature set is 6.8% above that observed when using the baseline set. The combined feature set outperforms the individual sets in all cases, the achieved accuracy being very close to that of the baseline feature set. This difference in accuracies is significant at the 0.05 level only in two cases (Ballroom and Unique). With regard to the datasets, results show accuracies in the area of 44.6 (rhythm feature set, Homburg) to 72.8% (combined feature set, GTZAN). The best performance of the rhythm features can be observed for the Ballroom dataset (67.7%), whereas the poorest performance can be found for the Homburg dataset. It should be noted that for the unbalanced datasets (i.e., all except GTZAN) it was observed from inspecting the confusion matrices that the achieved accuracy is mostly influenced by the most prominent class being classified correctly, whereas for the other classes, the performance is inferior but still in most cases above the prior of the respective class. An important result is the performance of the Ballroom dataset when using the annotated bars as texture window length (Table 2): It shows a clear improvement for the rhythm feature set alone and reaches 88.4% for the combined set.

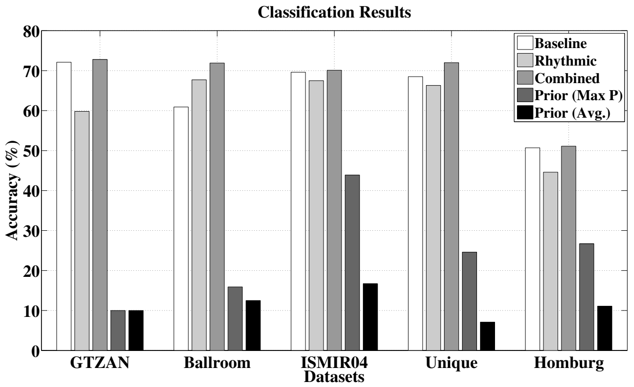

Figure 2: *Classification results after feature selection.*

Table 2: *Classification results (accuracy in %) for various settings and best classification results. For the Ballroom dataset, a second accuracy value is reported for the case of using the manual annotation for frame boundaries.*

Setting	GTZAN	Ballroom	ISMIR04	Unique	Homburg
Baseline	72.1	60.9	69.6	68.5	50.7
Rhythmic	59.3	67.3/76.7	67.5	66.3	44.6
Combined	72.8	71.9/88.4	70.1	72	51.1
P.-Max$P$	10	15.9	43.9	24.6	26.7
P.-Avg	10	12.5	16.7	7.1	11.1

Concerning feature selection, it can be seen in Table 3 that the best features resulting from the information-theoretical feature selection procedure comprise proportionally more features based on spectral (SF, SFL) or amplitude (RMS) novelty, whereas with respect to the subfeatures the image is not so clear — only the SD feature appears more consistently (4 times) in the first ranks. The worst features can be seen to be based on tonal (STPR, SPC) novelty functions, with subfeatures giving a clearer image, with Peak features such as the A1 being frequent in the ranking.

Those tendencies can also be partially observed in Fig. 3, where the feature groups were tested individually: the only novelty functions showing to provide slightly better results on its own for all datasets are the SFL, MFCC2, STPR and RMS, whereas other MFCC and SPC features show relatively lower performance on their own. However, the performance of all novelty functions appears to be relatively similar, except for the SFL, STPR and RMS novelty functions which seem to provide higher performance for all datasets. In the case of the subfeature groups, no specific feature seems to be standing out.

Table 3: *Best and worst features after feature selection. Abbreviation left of point denotes subfeature, otherwise novelty function.*

Rank	GTZAN	Ballroom	ISMIR04	Unique	Homburg
1	MD.RMS	P1.SF	MD.MFC2	SD.MFC1	SD.RMS
2	FL.RMS	A0.SFL	CD.MFC1	GM.SFL	SD.SPC3
3	GM.SFL	SD.SPC3	A0.SF	MD.MFC2	FL.SFL
568	A0.STPR	A0.STPR	SP.MFC3	A0.STPR	A1.MFC2
569	SP.MFC1	A1.MFC1	A1.RMS	A1.MFC3	A0.MFC2
570	A1.RMS	EN.MFC1	A0.STPR	A0.MFC3	A1.MFC1

Table 4: *Confusion matrix for Ballroom dataset, average accuracy: 67.3%. Accuracy and Prior are given in %.*

	Ch.	Ji.	Qu.	Ru.	Sa.	Ta.	Vw.	Wa.
Ch.	87	4	3	4	8	3	2	0
Ji.	9	40	1	1	6	2	1	0
Qu.	2	5	50	5	10	6	3	1
Ru.	10	0	3	62	0	5	2	16
Sa.	9	6	7	3	55	4	2	0
Ta.	3	0	9	2	2	58	7	5
Vw.	0	0	11	9	0	5	25	15
Wa.	0	0	2	12	0	1	2	93
Acc.	78	67	61	63	64	67	38	85
Pr.	15.9	8.6	11.7	14.0	12.3	12.3	9.3	15 .8

Table 5: *Confusion matrix for GTZAN dataset, average accuracy: 59.3%. Accuracy and Prior are given in %.*

	Bl.	Cl.	Co.	Di.	Hi.	Ja.	Me.	Po.	Re.	Ro.
Bl.	59	2	4	4	5	5	8	1	7	6
Cl.	2	79	1	1	0	11	2	0	1	5
Co.	11	3	56	4	2	7	3	1	1	13
Di.	3	0	4	56	6	3	7	8	6	10
Hi.	3	0	0	8	64	2	2	5	13	3
Ja.	7	13	4	1	0	64	4	1	3	3
Me.	2	2	1	8	1	1	72	2	0	12
Po.	7	2	3	9	5	4	1	42	10	16
Re.	6	1	3	6	14	4	0	4	59	3
Ro.	7	4	8	4	4	4	15	9	4	42
Acc.	59	79	56	56	64	64	72	42	59	42
Pr.	10	10	10	10	10	10	10	10	10	10

5. DISCUSSION

The results given in Table 2 provide strong support for the view that the presented rhythm features can perform in the same range as non-rhythmic baseline features. The overall better performance of the combined feature set is a consequence of using information related to both timbre and rhythm features, with significant improvement for two out of five datasets.

Concerning the datasets, the poor classification performance observed for the Homburg dataset could be an indication that the latter has a special genre or track selection, which cannot be predicted efficiently using our features. The very similar results observed for all feature sets for the ISMIR04 dataset is most probably due to its largely unbalanced character. The Ballroom dataset stands out as a good example where the rhythm features alone could offer a satisfactory performance. The results reported here are comparable to or better than those reported in studies using similar methods [16, 17, 18], but lie below those of newer studies employing more sophisticated features which depart from the beat histogram [55, 7, 10, 3, 56, 11, 57]. However, the simplicity of the beat histogram calculation and feature extraction and the possibility to assess features and feature groups individually are, in our view, advantages which have to be taken into account.

The use of bar boundaries as texture window boundaries for the Ballroom dataset gave encouraging results: A maximum accuracy of 76.7% was achieved, which is a notable improvement to the 67.7% with fixed-length segmentation. The result is close to the one reported by Gouyon et al. [18] using features extracted

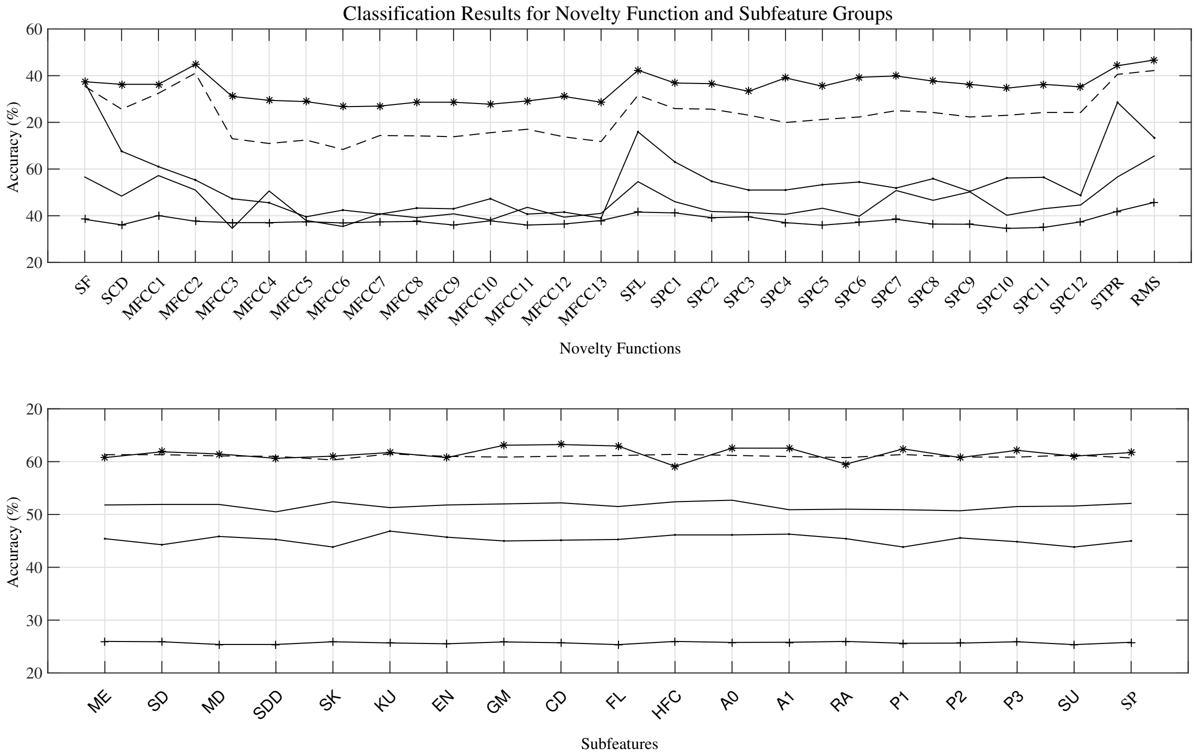

Figure 3: *Classification results for feature groups. Upper graphic shows Novelty Function, lower Subfeature groups. Datasets are denoted by the following line styles: "-" (GTZAN), ".-" (Ballroom), "*-" (ISMIR04), "–" (Unique), "+-" (Homburg).*

from the beat histogram. This suggests that the application of prior knowledge regarding the "real" boundaries of the musical surface can help to considerably improve the accuracy of rhythm-based genre classification.

The misclassifications (reported in Tables 4 and 5) show that genre confusions are those that could be expected when using features capturing the rhythmic character of the pieces, thus making it difficult to distinguish between genres containing similar patterns: for the Ballroom dataset, the most prominent misclassifications take place between Rumba/Waltz, Quickstep/Samba and Waltz/Viennese Waltz. For the GTZAN dataset, Classical is confused with Jazz, Country with Rock, Hiphop with Reggae and Pop with Rock and Metal. Those results indicate that the multiple novelty-based beat histogram features indeed capture the specific rhythmic properties of the genres and the regularities of their periodicities, pointing towards the suitability of those features for the extraction of general rhythmic properties.

With regards to the feature selection, it was shown that mainly novelty functions which are based on amplitude or spectral shape changes in the signal give the best results, a result which could be confirmed both from the standard feature selection procedure, as well as from the feature subset-based selection. Those results can be due to important turning points of a track in most popular western music being mostly mediated through loudness, energy or spectral form (for example, in the case of many instruments playing together in a bar change or a new instrument entering the scene) which leads to those features possessing more salient novelty functions out of which more qualitative beat histogram features can be extracted. Another reason could be the role of those novelty functions in stressing the basic metric positions in most of western popular music. Concerning the subfeatures and the results of the information-theoretical feature selection (Table 3), the higher performance of simple statistics such as the SD of the beat histogram attest to them having more discriminative power, possibly because they express very basic statistical tendencies of the periodicity distribution. Furthermore, higher-level features (such as the P1, which is a good estimate of the excerpts' tempo, already shown to be important in rhythm-based genre classification [18]), also possess discriminative power since their value is an important indicator of a genre character (e.g. dance vs. classical music). However, since no such tendencies are observed in the feature group selection (Fig. 3), it can be deducted that no specific subfeature on the beat histogram bears special discriminative power, but it is rather the combination with a salient novelty function which allows for better results.

6. CONCLUSIONS

The work presented in this paper focuses on the creation of novel features for rhythm-based musical genre classification. The difference in comparison to previous studies in the field [16, 17, 18], using the signal amplitude envelope only, is the use of the temporal trajectory of other signal quantities such as SF as the novelty function for the calculation of the beat histogram. We showed that performance using these beat histogram features is higher or in a similar range than related work using periodicity histograms. It has also been shown that specific novelty functions relating to amplitude or spectral shape are among the most informative when analyzed with a periodicity representation method. Finally, we showed the positive impact of manual bar-boundary annotation for the extraction of rhythm features on classification performance.

There are many more features which can be considered as novelty functions [38, 43], as well as possible subfeatures on the beat histogram (such as MFCCs, presented in [18]) and other methods for the periodicity representation calculation [19, 17, 18]. Future goals include an even more extensive and detailed feature selection, identifying the features or feature groups which are informative with respect to specific genres (in order to associate specific novelty functions with relevance to specific genres) and a test of the feature robustness against signal degradations. Those subfeatures and novelty functions could be then suitable for future use in more specific rhythm description tasks. Furthermore, the investigation of optimal parameter settings for feature extraction and classification, the utilization of other classification methods and performance evaluation measures, as well as the usage of other methods for feature aggregation are other possible research directions. The high accuracy achieved especially for the Ballroom dataset indicates the suitability of the descriptors for further application and points to the importance of bar-boundary annotation (which can also be performed automatically) for rhythm features.

While the features for beat histogram calculation have been evaluated only in the context of genre classification, we believe that they will prove useful in other tasks as well. Future research will concentrate on adjusting and using rhythm content features for MIR tasks such as audio similarity and mood recognition. Preliminary research has also been undertaken concerning the use of novelty functions of specific instruments (e.g. Drums) extracted through Non-Negative Matrix Factorization (NMF) [58], or the application of the features to other signals, such as speech [59]. Their application in a task of automatic spoken language identification based on the rhythmic elements of speech has shown promising results and points to further research directions.

7. REFERENCES

[1] Enric Guaus i Termens, “New approaches for rhythmic description of audio signals,” Tech. Rep., Universitat Pompeu Fabra, Music Technology Group, 2004.

[2] Fabien Gouyon and Simon Dixon, “A review of automatic rhythm description systems,” *Computer Music Journal*, vol. 29, no. 1, pp. 34–35, 2005.

[3] Geoffroy Peeters, “Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1242–1252, 2011.

[4] Tlacael Miguel Esparza, Juan Pablo Bello, and Eric J Humphrey, “From genre classification to rhythm similarity: Computational and musicological insights,” *Journal of New Music Research*, , no. ahead-of-print, pp. 1–19, 2014.

[5] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek, “Automatic genre classification of music content: a survey,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, March 2006.

[6] Enric Guaus i Termens, *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers.*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, 2009.

[7] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer, “On rhythm and general music similarity.,” in *ISMIR*, 2009, pp. 525–530.

[8] François Pachet, Daniel Cazaly, et al., “A taxonomy of musical genres,” in *Proceedings of the Conference on Content-Based Multimedia Information Access*, 2000.

[9] Jean-Julien Aucouturier and Francois Pachet, “Representing musical genre: A state of the art,” *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[10] Andre Holzapfel, Arthur Flexer, and Gerhard Widmer, “Improving tempo-sensitive and tempo-robust descriptors for rhythmic similarity,” in *Proceedings of the 8th Sound and Music Computing Conference*, 2011.

[11] Aggelos Pikrakis, “A deep learning approach to rhythm modelling with applications,” in *6th International Workshop on Machine Learning and Music (MML13)*, 2013.

[12] Grosvenor Cooper, *The rhythmic structure of music*, vol. 118, University of Chicago Press, 1963.

[13] Paul Fraisse, “Rhythm and tempo,” in *The psychology of music*, Diana Deutsch, Ed., Series in Cognition and Perception, chapter 6. Academic Press, 1982.

[14] Fred Lerdahl and Ray S Jackendoff, *A generative theory of tonal music*, MIT press, 1983.

[15] Joel Lester, *The rhythms of tonal music*, Pendragon Press, 1986.

[16] George Tzanetakis and Perry Cook, “Musical genre classification of audio signals,” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[17] Juan José Burred and Alexander Lerch, “A hierarchical approach to automatic musical genre classification,” in *DAFX*, 2003.

[18] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer, “Evaluating rhythmic descriptors for musical genre classification,” in *AES*, 2004.

[19] Eric D Scheirer, “Tempo and beat analysis of acoustic musical signals,” *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[20] Elias Pampalk, Simon Dixon, and Gerhard Widmer, “Exploring music collections by browsing different views,” *Computer Music Journal*, vol. 28, no. 2, pp. 49–62, 2004.

[21] Simon Dixon, Fabien Gouyon, Gerhard Widmer, et al., “Towards characterisation of music via rhythmic patterns,” in *ISMIR*, 2004.

[22] Andre Holzapfel and Yannis Stylianou, “Rhythmic similarity of music based on dynamic periodicity warping,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2217–2220.

[23] Simon Dixon, Elias Pampalk, and Gerhard Widmer, “Classification of dance music by periodicity patterns,” in *ISMIR*, 2003.

[24] Jonathan Foote and Shingo Uchihashi, “The beat spectrum: A new approach to rhythm analysis,” in *ICME*, 2001.

[25] Kristopher West and Stephen Cox, "Features and classifiers for the automatic classification of musical audio signals," in *ISMIR*, 2004.

[26] Elias Pampalk, Arthur Flexer, Gerhard Widmer, et al., "Improvements of audio-based music similarity and genre classification," in *ISMIR*. London, UK, 2005, vol. 5, pp. 634–637.

[27] Geoffroy Peeters and Helene Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1754–1769, 2011.

[28] Bob L Sturm, Corey Kereliuk, and Aggelos Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in *Cognitive Information Processing (CIP), 2014 4th International Workshop on*. IEEE, 2014, pp. 1–6.

[29] George Tzanetakis, Georg Essl, and Perry Cook, "Human perception and computer extraction of musical beat strength," in *Proc. DAFx*, 2002, vol. 2.

[30] Bob L Sturm, "An analysis of the gtzan music genre dataset," in *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM, 2012, pp. 7–12.

[31] Bob L Sturm, "The gtzan dataset: Its contents, its faults, their affects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[32] B Sturm, "A simple method to determine if a music information retrieval system is a "horse"," *IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 16, NO. 6, OCTOBER 2014*, 2014.

[33] H Christopher Longuet-Higgins and Christopher S Lee, "The perception of musical rhythms," *Perception*, vol. 11, no. 2, pp. 115–128, 1982.

[34] Richard Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music Perception*, pp. 409–464, 1994.

[35] David Temperley, *The cognition of basic musical structures*, MIT press, 2004.

[36] Justin London, *Hearing in time*, Oxford University Press, 2012.

[37] Stephen Hainsworth and Malcolm Macleod, "Onset detection in musical audio signals," in *ICMC*, 2003.

[38] Juan P. Bello, Chris Duxbury, Mike Davies, and Mark Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553–556, 2004.

[39] Axel Roebel, "Onset detection in polyphonic signals by means of transient peak classification," in *ISMIR*, 2005.

[40] Anssi Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999.* IEEE, 1999, vol. 6, IEEE.

[41] Chris Duxbury, Mark Sandler, and Mike Davies, "A hybrid approach to musical note onset detection," in *Proc. Digital Audio Effects Conf.(DAFX,í02)*, 2002, pp. 33–38.

[42] Chris Duxbury, Juan Pablo Bello, Mike Davies, Mark Sandler, et al., "Complex domain onset detection for musical signals," in *Proc. Digital Audio Effects Workshop (DAFx)*, 2003, pp. 6–9.

[43] Juan P. Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[44] Alexander Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*, John Wiley & Sons, 2012.

[45] Vladimir Vapnik, *The nature of statistical learning theory*, Springer, 2000.

[46] Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, vol. 28, no. 2, pp. 63–76, 2004.

[47] Klaus Seyerlehner, Gerhard Widmer, and Dominik Schnitzer, "From rhythm patterns to perceived tempo," in *ISMIR*, 2007.

[48] Helge Homburg, Ingo Mierswa, Bülent Möller, Katharina Morik, and Michael Wurst, "A benchmark dataset for audio classification and clustering," in *ISMIR*, 2005.

[49] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[50] Hanchuan Peng, Fulmi Long, and Chris Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2005.

[51] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[52] Ron Kohavi and George H John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[53] Florian Krebs, Sebastian Böck, and Gerhard Widmer, "Rhythmic pattern modeling for beat and downbeat tracking in musical audio.," in *ISMIR*, 2013.

[54] Giles M Foody, "Thematic map comparison," *Photogrammetric Engineering & Remote Sensing*, vol. 70, no. 5, pp. 627–633, 2004.

[55] Klaus Seyerlehner, Markus Schedl, Tim Pohle, and Peter Knees, "Using block-level features for genre classification, tag classification and music similarity estimation," *MIREX 2010*, 2010.

[56] Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R Arce, "Music genre classification using locality preserving non-negative tensor factorization and sparse representations.," in *ISMIR*, 2009.

[57] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce, "Music genre classification via joint sparse low-rank representation of audio features," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1905–1917, 2014.

[58] Athanasios Lykartsis, Chih-Wei Wu, and Alexander Lerch, "Beat histogram features from nmf-based novelty functions for music classification," in *ISMIR*, 2015.

[59] Athanasios Lykartsis and Stefan Weinzierl, "Using the beat histogram for speech rhythm description and language identification," in *INTERSPEECH*, 2015.