Audio Engineering Society

Convention Paper 7268

Presented at the 123rd Convention
2007 October 5–8 New York, NY

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Towards Textual Annotation of Rhythmic Style in Electronic Dance Music

Kurt Jacobson[1], Matthew Davies[1], and Mark Sandler[1]

[1] *Centre for Digital Music, Queen Mary University of London, Mile End Road, E1 4NS, UK*

Correspondence should be addressed to Kurt Jacobson (kurt.jacobson@elec.qmul.ac.ukl)

ABSTRACT

Music information retrieval encompasses a complex and diverse set of problems. Some recent work has focused on automatic textual annotation of audio data, paralleling work in image retrieval. Here we take a narrower approach to the automatic textual annotation of music signals and focus on rhythmic style. Training data for rhythmic styles are derived from simple, precisely labeled drum loops intended for content creation. These loops are already textually annotated with the rhythmic style they represent. The training loops are then compared against a database of music content to apply textual annotations of rhythmic style to unheard music signals. Three distinct methods of rhythmic analysis are explored. These methods are tested on a small collection of electronic dance music resulting in a labeling accuracy of 73%.

1. INTRODUCTION

Music is a very complex medium. A music signal can be recorded and perfectly reconstructed using digital technology. However, the essence of what is actually communicated by that music signal is much more elusive. This makes organizing and accessing large stores of music content in a meaningful and intuitive way very difficult. A variety of approaches to music retrieval have been proposed including query-by-humming, query-by-example, and qeury-by-style or genre. However, the most natural way for a listener to describe a piece of music or to formulate a music query is with words. This combined with the fact that text-based query is a mature discipline with proven methods makes the automatic textual annotation of music signals an attractive goal.

While different listeners may use very different words to describe the same piece of music, there are some semantic descriptions of music that are generally agreed upon. For example, there is at least some level of consensus regarding descriptions such as genre, instrumentation, or rhythmic style. Previous work has attempted to automatically annotate music signals for the purpose of a query-by-text music retrieval system [1, 2, 3]. These systems attempt to annotate music signals in a general sense, not focusing on any specific aspect of the music.

However, there are several dimensions of musicality inherent to a music signal. These include melody, harmony, timbre/instrumentation, and rhythm among others [4]. We purpose a 'divide-and-conquer' approach to automatic textual annotation of music signals, dealing with each aspect of musicality individually. This work deals with only the rhythm of a music signal. There is a relatively high level of agreement with respect to the textual descriptions of rhythmic styles. For example, rhythmic styles tend to be fairly well-defined in musicology. There is little debate regarding what rhythms constitute a "waltz" versus a "samba". Our aim is to exploit this fact for automatic textual annotation.

Section 2 describes some previous work on rhythmic analysis, section 3 describes the methods used in the current work, section 4 discusses the results and section 5 suggests future work.

2. RELATED WORK

In addition to work on automatic textual descriptions of music, there is a fairly large body of work related to audio-based rhythmic analysis. Foote et. al. purposed a method for rhythmic analysis and similarity based on features of a self-similarity matrix derived from audio frames [5]. Dixon et. al. developed a method for using bar-length pattern features from audio and applied these features to genre classification [6]. A robust method for automatically extracting bar-length patterns from audio is described by Davies [7]. A study exploring various rhythmic descriptors and their effectiveness can be found in [8].

3. METHODS

As a first step towards the textual annotation of rhythmic style using drum loops as source data, we treat the problem as a classification task where the training data is restricted to drum loop material. Unheard music pieces from a small collection are to be classified by their rhythmic style based on this training data alone.

3.1. Source Material

A collection of 30 drum loops from three electronic dance music genres are used as training data. The collection includes 10 loops representing Hip-Hop rhythmic style, 10 loops representing a four-on-the-floor House rhythmic style, and 10 loops representing a Drum-and-Bass (DnB) rhythmic style. For each drum loop, 8 measures are used. The tempos range from 88 bpm to 175 bpm. The drum loops were obtained from the popular sound synthesis and sequencing package Reason [9].

A small collection of electronic dance music containing 140 pieces was used for testing material. The rhythmic style of each piece in the collection is annotated manually to provide ground truth data. The collection contains 46 pieces with a Hip-Hop rhythm, 43 pieces with a House rhythm, and 51 pieces with a DnB rhythm.

3.2. Analysis

Three distinct analysis methods are applied to the classification task. These include beat spectrum analysis, bar-length pattern analysis, and cross-correlation tempo estimation.

3.2.1. Beat Spectrum Approach

The *beat spectrum* is calculated for each training drum loop as proposed by Foote et al [5]. The signals are divided into overlapping frames with a length of 11ms and an FFT is performed on each frame. The cosine distance between magnitude spectra is used to compute similarity between frames for every frame within a 4 second segment. This results in a square self-similarity matrix. The diagonals of the self-similarity matrix are summed resulting in a vector describing the rhythm and tempo of the loop - the beat spectrum. The result is a set of feature vectors describing tempo and rhythm that are associated with a particular textual annotation for rhythmic style.

The same beat spectrum extraction is performed on the test set of music signals. However, the test music signals are significantly longer than the rhythm loops. Therefore, the beat spectrum is calculated for ten different 4 second segments in the test music signal spaced evenly in time across the song. The resulting beat spectrum vectors are clustered using a k-means algorithm (with k=3). The mean of the largest cluster is used as the representative beat spectrum vector for that test music signal. This helps to exclude segments that might contain a break in the rhythm not characteristic of the entire piece.

For each music signal in the test collection, a comparison is made to each drum loop in the training set. The cosine distance measure is used to measure the similarity between pairs of beat spectra. The test music signal is then assigned the label of the closest matching drum loop.

3.2.2. Bar-Length Pattern Classification

The bar-length pattern approach is derived from the method employed by Dixon et al [6] for the classification of ballroom dance music. The feature used to characterise the rhythmic style of the input musical signal is a continuous bar-length pattern which emphasizes the locations of note onsets and their relative metrical strengths. The continuous feature used is the complex spectral difference onset detection function [10]. The principal difference between our approach and that of Dixon et al is the use of a fully automatic method for extracting the bar pattern [7]. Dixon et al rely on manually annotated data to derive the bar patterns through semi-automatic analysis.

The automatic extraction of the bar pattern requires several steps, the first of which is the generation of the onset detection function across the entire length of the musical excerpt. This signal is then passed to a beat tracker which returns beat locations corresponding to periodic, strong peaks in the onset detection function.

Given the sequence beat times, the bar boundaries are then extracted by measuring the dissimilarity between beat-synchronous spectral frames. The beat frames leading to consistent spectral difference are taken to indicate the first beat of each bar (i.e. the downbeat). Since all our training and test excerpts have 4/4 time-signature, we need not estimate the time-signature automatically.

To identify the representative bar-length pattern for a given music signal we extract each individual bar from the onset detection function and resample it to have constant length (set to 144 detection function samples, with resolution 11.6ms per sample). We then follow Dixon et al [6] and cluster the extracted bars using k-means (with k=3). The predominant bar pattern is taken as the temporal mean of the largest cluster. Further details on the extraction of the rhythmic information can be found in [7].

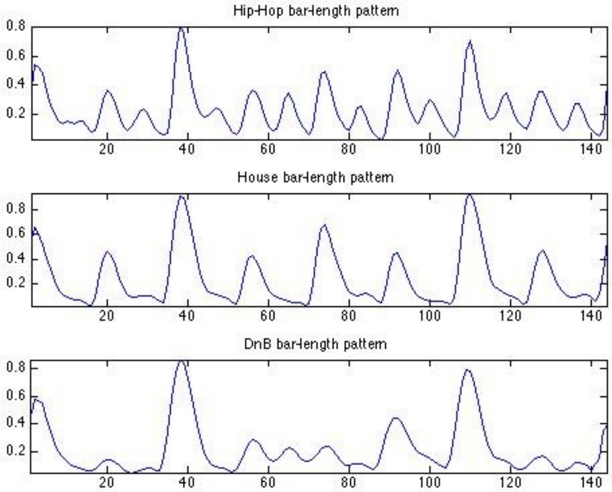

Fig. 1: Averaged bar-length patterns for hip-hop, house, and dnb training loops.

To classify the rhythmic style of a given bar length pattern $\gamma(m)$ within our test set, we compare it to the predominant bar pattern $\Gamma_p(m)$ for each class p in our training set. The bar patterns for each rhythmic style are calculated as the temporal mean of the bar patterns for each training excerpt. We provide two methods for rhythmic style classification. The first $D_1(p)$ is calculated as the Euclidean distance between the test pattern and training patterns,

$$D_1(p) = \sum_{m=1}^{M} |\gamma(m) - \Gamma_p(m)|^2. \qquad p = 1, \ldots, P. \tag{1}$$

where $P = 3$. The second measure $D_2(p)$ scales each D_1 by the distance between the estimated tempo τ of the test bar pattern $\gamma(m)$ with the nominal tempo of each class $T_p = \{90, 120, 160\}$,

$$D_2(p) = D_1(p)\left(1 + \left|1 - \frac{\tau}{T_p}\right|\right) \qquad p = 1, \ldots, P. \tag{2}$$

where tempo is measured in beats per minute. The decision over rhythmic style for $D_1(p)$ and $D_2(p)$ is found as the p which gives the minimum distance.

3.2.3. Cross-Correlation Tempo Estimation

Like most tempo estimation methods, the tempo calculation used in the bar-length pattern approach is prone to double-time and half-time errors. The tempo of a track that should be labeled “drum and bass” would often be calculated to be 90 bpm when the actual tempo was 180 bpm. Similar errors occur on the training drum loops when the tempo is calculated automatically.

This problem inspires a new approach to tempo estimation. Using the representative bar-length patterns calculated for each classification label and the automatic onset detection function of an unknown query music signal, a more accurate tempo estimation is performed. The bar-length patterns are stretched in time to cover a range of possible tempos. Tempos from 80 bpm to 195 bpm are covered at 1 bpm intervals. For each stretch, a cross-correlation between the onset function and the bar-length pattern is calculated. The stretch that results in the highest cross-correlation value corresponds to the estimated tempo of the query music signal.

Because we have three bar-length patterns (one for each label), this results in three tempo estimates. A majority-rules approach is used to arrive at one tempo value for the music signal in question. Then, labels are applied using the following rule set:

$$\text{Genre} = \begin{cases} \text{Hip-Hop} & : \text{tempo} < 102 \\ \text{House} & : 102 <= \text{tempo} < 138 \\ \text{DnB} & : \text{tempo} >= 138 \end{cases} \tag{3}$$

4. RESULTS

The test set used in this study contains classification boundaries that are highly dependent on tempo. Therefore, it is not surprising that the best classification results are achieved using the cross-correlation tempo estimation and classifying along tempo boundaries. This results in 72.8% classification accuracy. The beat spectrum approach achieves 68.6% accuracy and the bar-length pattern approach, when combined with tempo estimations, achieves up to 68.6% accuracy. The baseline classification for the test collection was 36.4%.

The beat spectrum approach seems to mostly confuse hip-hop rhythmic style with drum and bass rhythmic style. This is probably related to the double-time / half-time errors that are so common in rhythmic analysis. The confusion matrix for the beat spectrum approach is shown in Table 1.

Label	HipHop	House	DnB
HipHop	23	0	23
House	4	29	10
DnB	7	0	44

Table 1: Beat spectrum confusion matrix

Label	HipHop	House	DnB
HipHop	26	3	17
House	4	32	7
DnB	8	5	38

Table 2: Bar-length pattern confusion matrix

With respect to the bar-length pattern approach, the best results are obtained using the Euclidean-distance-scaled-by-tempo classification scheme which results in a classification accuracy of 67.1%. However, using the calculated tempo alone results in a classification accuracy of 62.1%. Without tempo scaling, classification accuracy was just below baseline at 35.3%. When using the tempo estimation from the cross-correlation approach to scale the Euclidean distance, a classification accuracy of 68.6% is achieved. The confusion matrix for the bar-length pattern approach is shown in Table 2.

After visually inspecting the average bar-length patterns from the training data in Figure 1, it is apparent that the bar-length patterns associated with each label are very similar. In each rhythmic style, there are very strong onsets on the 2 and the 4 beats - corresponding to 36/144 and 108/144 respectively in the normalized bar-length pattern. This is because each rhythmic style includes strong snare hits on the up beats. Similarly, in each bar-length pattern there is regular onset energy corresponding to 1/8th and 1/16th notes. Ignoring tempo information, the three styles explored in this work appear very similar.

The best results were obtained using the cross-correlation tempo estimation and a simple decision tree approach to classification. This approach works well for the current data set, which is largely divided along tempo boundaries. Of course, this assumption cannot be made when trying to extend textual anno-

Label	HipHop	House	DnB
HipHop	27	4	15
House	4	33	6
DnB	3	6	42

Table 3: Cross-correlation tempo confusion matrix

tation to the general case. The confusion matrix for the cross-correlation tempo estimation classification is shown in Table 3.

5. FUTURE WORK

Given the small size of the test set, the current results are not as encouraging as those reported in [6]. This is partially due to the inherent rhythmic similarity across rhythmic labels in the test set. Figure 1 illustrates this similarity and the results indicate the importance of tempo in rhythmic analysis. However, there are many opportunities for improvement over the current approach.

The three approaches used in this study could possibly be combined to increase performance. Alternative rhythmic analysis approaches might be more appropriate for the task of automated textual annotation of rhythmic style. More advanced approaches related to drum track transcription have been proposed [11]. Additional classification schemes should be explored as well. Hidden Markov models (HMM) or support vector machines (SVM) could be applied to the training and classification tasks.

A much larger training set of drum loops could be used. However, the appropriate textual label to associate with a particular loop can be ambiguous. Even given the three example labels used in this study there is some ambiguity. For example, a 'house' label would be applied to a 'trance' track because both have a similar rhythmic template and tempo. However, this result might not be desirable for indexing and searching through collections of dance music. Ultimately, a much larger testing set should also be used, including more genres and more diversity in rhythmic style.

In practice, a system for automatic textual annotation of rhythmic style should be able to make a decision regarding the appropriateness of any automatically assigned annotation, perhaps deciding no label can be applied. This would be a critical feature of any practical automatic textual annotation system but was not considered here.

6. ACKNOWLEDGMENTS

This research was supported by EPSRC grant EP/E017614/1 (Online Music Recognition And Searching).

7. REFERENCES

[1] D. Turnbull, L. Barrington, and G. Lanckriet, "Modeling Music and Words Using a Multi-class Naive Bayes Approach," in *Proceedings of 7th ISMIR*, Victoria, Cannada, 2006.

[2] B. Whitman and R. Rifkin, "Musical Query-by-Description as a Multiclass Learning Problem," *IEEE Multimedia Signal Processing*, 2002.

[3] B. Whitman, "Learning the Meaning of Music", M.I.T. PhD Thesis, 2005.

[4] P. Herrera et. al., "SIMAC: Semantic Interaction with Music Audio Contents," in *Proceedings of 5th ISMIR*, Barcelona, Spain, 2004.

[5] J. Foote, M. Cooper, and U. Nam, "Audio Retrieval by Rhythmic Similarity," in *Proceedings of 3rd ISMIR*, 2002.

[6] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *Proceedings of 5th ISMIR*, Barcelona, Spain, pp. 509–517, 2004.

[7] M. Davies, "Towards Automatic Rhythmic Accompaniment", Queen Mary PhD thesis, 2007.

[8] F. Gouyon, et al, "Evaluating Rhythmic Descriptors for Musical Genre Classification", AES 25th International Conference, 2004.

[9] http://www.propellerheads.se/

[10] J. Bello, C. Duxbury, M. Davies, and M. 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.

[11] O. Gillet and G. Richard, "Drum Track Transcription of Polyphonic Music Using Noise Subspace Projection," in *Proceedings of 6th ISMIR*, London, UK, 2005.