Preprocessing methods for rhythmic mid-level features

Christian Dittmar, Matthias Gruhne, Daniel Gaertner
Fraunhofer Institute for Digital Media Technology, 98693 Ilmenau, Germany,
Email:{dmr,ghe,gtr}@idmt.fraunhofer.de

Abstract

Rhythmic mid-level features are an important prerequisite in Music Information Retrieval. They can be deployed to describe the rhythmical gist of songs for various tasks, such as genre classification and music similarity search. Although different computation strategies have been proposed in the literature, rhythmic mid-level features commonly represent the most salient rhythmic periodicities in a music signal. Major shortcomings of rhythmic mid-level features are their dependency on the actual tempo of the songs and their susceptibility to degradation caused by interference with non-rhythmic signal components, such as melodic-sustained instruments. This publication describes preprocessing strategies which can be applied to rhythmic mid-level features. The problem of interference with melodic instruments is addressed by emphasizing the rhythmic content via a signal decomposition related to Drum Transcription. Tempo changes are tackled by logarithmic re-sampling of the features' lag axis. An evaluation of the proposed methods is conducted via genre classification using both artificial and real-world music signals. The evaluation results are presented and discussed accordingly.

Introduction

During recent years the scientific and commercial interest in automatic methods for music search and recommendation has significantly increased. Stimulated by the ever-growing availability and size of digital music collections, methods from the field of Music Information Retrieval (MIR) pose increasingly important means to enable convenient exploration of large music catalogs. Evidently, commercial entities like online music shops have realized that there is a need to enrich the virtual goods they sell with as much additional information as possible. Automatically classified categorizations in terms of genre and mood are common examples of such metadata. Efficient description of the rhythmic gist inherent in music pieces is one important pre-requisite for such tasks. A number of different approaches towards the extraction of so-called rhythmic mid-level features are described in the literature [3], [6], [12]. The majority of approaches utilizes an auto-correlation function (ACF) of temporal envelopes taken from a time-frequency representation of the audio signal. The resulting rhythmic mid-level feature is called Beat Histogram [12] or Beat Spectrum [4]. The most salient peaks of such a feature represent the strongest periodicities in the music. They are arranged on the so-called lag-axis and can be directly converted into tempo hypotheses, expressed in beats per minute (BPM). Unfortunately, these features show two negative properties, that constrict direct usage in terms of distance computation or statistical modeling.

Tempo dependency: Beat spectra are dependent on the actual tempo of the song. That means that the shape of the ACF and thus also the lag-axis positions of the peaks are stretched in case of a tempo decrease and compressed in case of a tempo increase. Two songs with exactly the same beat played at different tempi will exhibit low similarity when computing a simple distance measure between the two Beat Spectra. In [4] it is concluded, that the distance between the rhythmic mid-level feature vectors increases linearly with the tempo difference. In [12] statistical measures are derived from the mid-level representation: the relative amplitude from the first two peaks, the ratio of the amplitude of the second peak divided by the amplitude of the first peak, the lag-period of the first and the second peak and the overall sum of the vector. Gouyon [6] computes a slightly different feature consisting of: mean-value, geometric mean, total energy, centroid, flatness, skewness and high-frequency content. Burred [1] uses a similar feature-set but extends it by the standard deviation, mean of the derivative, the standard deviation of the derivative, and the entropy. However, the classification results with some of the described features are still showing tempo dependency. Some features introduced by Tzanetakis refer to the absolute lag-position of the first and second local maximum, which is clearly tempo dependent. Gouyon and Burred used statistical moments, which are likely to change significantly, when the tempo and therewith the stretching of the corresponding Beat Spectra changes.

Periodicity degradation: Beat Spectra tend to become more noisy and flat if the rhythmic components of the music signal are interfering with competing signals that are not impulsive and necessarily supporting the percussive rhythm. For example, melodic-sustained instruments as well as more dynamic signals like singing voice clearly decrease the ACF's peak magnitudes (i.e., periodicities). This makes it difficult to extract the real rhythmic signal. Therefore a splitting into percussive vs. sustained components as proposed in [13] and later picked up in [8] are probably suitable to attenuate the negative influences of melodic instruments.

Proposed preprocessing

This publication essentially describes two different preprocessing steps that can be integrated into the signal-processing chain before the computation of statistical descriptors of the Beat Spectrum. For extracting the raw

Beat Spectrum, the method described in [3] is deployed. The Beat Spectra are given by the ACF of a weighted and differentiated sum of the single amplitude envelopes of the Audio Spectrum Envelope (ASE). The accumulation incorporates a weighting of the single band envelopes with the so-called Percussiveness measure, introduced in [13]. This grants an emphasis of rhythmically significant signal components. For the ACF computation, excerpts of 500 frames were chosen, which corresponds to 5 seconds in music, given a low-level hop-size of 10 milliseconds. This size constitutes a trade-off between the length of at least two repeating patterns and the ability to track abrupt tempo changes sometimes encountered in real-world music.

Tempo independence by logarithmic re-sampling: As stated before, the ACF-based Beat Spectrum is not tempo independent, since the shape of the mid-level feature vector is compressed or expanded depending on the tempo of the rhythm. The compression of such a vector can be interpreted as multiplication of a time-stretching factor f with the underlying pattern signal $\mathbf{c}(\tau')$. The observed feature vector can therefore be described as $\mathbf{c}(\tau) = \mathbf{c}(\tau' \cdot f)$. In order to obtain a tempo invariant Beat Spectrum $\mathbf{c}(\tau')$, the stretch factor f needs to be known, but its automatic estimation might be unreliable. One option for solving this issue is logarithmic re-sampling of the lag-axis. By applying the logarithm to an arbitrary function, multiplicative terms are transformed to additive terms. Transferring this theorem to the lag-axis of the Beat Spectrum $\mathbf{c}(\tau)$ yields the equation $\mathbf{c}(\log(\tau' \cdot f)) = \mathbf{c}(\log(f) + \log(\tau'))$. The original rhythmic Beat Spectrum is transformed by applying the logarithm on the function τ, which leads to the two additive parts: the logarithmized stretch factor $\log(f)$ and a tempo invariant feature vector $\log(\tau')$. For the logarithmic re-sampling, a new variable is estimated by $\tau_{log} = \frac{\log(\tau) \cdot max(\tau)}{\log(max(\tau))}$. Distorting the original Beat Spectrum $\mathbf{c}(\tau)$ from argument τ to argument τ_{log} results in a new mid-level feature vector with logarithmized lag-axis. Since τ_{log} consists of non-integer values, an application of this argument to the original rhythmic mid-level feature vector requires an interpolation. In order to obtain the tempo independent part, the logarithmized stretch factor $\log(f)$ needs to be removed from the feature vector $\mathbf{c}(\tau_{log})$. The goal is therefore to find the point, where the logarithmized stretch factor f ends in order to estimate a new vector $\mathbf{c}'(\tau_{log})$ starting from this point $\log(f)$ to the end of $\mathbf{c}(\tau_{log})$. By inspecting a large number of the logarithmically re-sampled Beat Spectra it can be observed, that all vectors consist of a large decaying slope towards a first local minimum, whose absolute position depends on the tempo of the music. That slope represents the first maximum lobe of the ACF. The first lobe is always the highest and does not carry any rhythmic information. However, the successive minimum appears to be the point right ahead of the logarithmized beat period. Since some feature vectors can also contain smaller local minima along the first slope towards the significant local minimum, it is not advisable to search for the very first local minimum. Instead, two different smoothed functions $\mathbf{l_1}(\tau_{log}), \mathbf{l_2}(\tau_{log})$ are estimated from the original curve $\mathbf{c}(\tau_{log})$, using a running average method, both with a different length of this running average smoothing S. The point τ_{log}, where both curves intersect is chosen as the end point of the logarithmized stretch factor $\log(f)$ and the beginning point of the tempo independent Beat Spectrum $\mathbf{c}'(\tau_{log})$. The tempo independent part contains a different number of elements, because the vector size of the original rhythmic midlevel feature is constant and the tempo dependent part has a different length, depending on the tempo. Therefore, in the experiments conducted for this publication, the resulting vector has been shortened to 250 elements. Sure enough, the stretch factor could also be found in the original Beat Spectrum and the transformation could be done in the original lag domain. But it should be noted, that the logarithmically spaced lag-axis inherently delivers some useful properties. The logarithmic spacing grants more vector elements to the areas that represent the tatum and beat [14], whereas the area representing the bar periodicity is strongly compressed. Thus, the periodicities corresponding to faster rhythmic levels are emphasized in contrast to the long term periodicities. This is especially important for a distance computation where all vector elements are treated equally.

Periodicity preservation by Drum Transcription: In [8], it is described how to derive amplitude envelopes that essentially capture the most important percussive instruments and suppress other melodic instruments. This is achieved by grabbing intermediate results available after the first two processing stages of the Drum Transcription algorithm detailed in [2]. The first stage consists of the collection of candidate onset times and their corresponding onset spectra. The second stage deploys higher order statistical computations in order to estimate frequency and amplitude bases of the involved percussive instruments. The unwrapped phase spectrogram $\mathbf{\Phi}$ and the magnitude spectrogram $\mathbf{X}$ of the music signal are derived by means of a conventional Short-Term Fourier Transform (STFT). The envelopes in each spectral bin are differentiated and Half-wave rectification is applied to derive a non-negative difference-spectrogram $\hat{\mathbf{X}}$. The detection of transient onsets $\mathbf{t}$ is conducted by means of peak picking in an onset-detection function derived by accumulating all bins of $\hat{\mathbf{X}}$ and smoothing the resulting vector. The main concept of the further process is the storage of one frame of $\hat{\mathbf{X}}$ at the time of the onset. Principal Component Analysis (PCA) is applied in order to compact the set of collected spectra $\hat{\mathbf{X}}_\mathbf{t}$ to a limited number of decorrelated and variance normalized (whitened) [9] principal components. The whitened components $\tilde{\mathbf{X}}$ are subsequently subjected to Non-Negative Independent Component Analysis (NN-ICA) [11]. The constraints for the NN-ICA model are that the original source spectra must be positive and well grounded and they must be to some extent linearly independent. The first requirement is always fulfilled because the vectors are subsets of the differentiated and half-wave rectified magnitude-spectrogram $\hat{\mathbf{X}}$ that does not contain any values lower than zero, but certainly some values at zero. The second constraint is taken into account when the spectra collected at onset times are regarded as linear superposition of a small set of original source-spectra characterizing the involved instruments. A useful property of percussive instruments is their nearly invariant overtone structure [2] as opposed to pitched sounds that constantly change with the melodic progression. This assumption holds up well in the majority of the cases, allowing to separate $\tilde{\mathbf{X}}$ according to 1,

$$\mathbf{F} = \mathbf{A} \cdot \tilde{\mathbf{X}} \tag{1}$$

where $\mathbf{A}$ denotes the unmixing matrix estimated by the NN-ICA, which does actually separate the individual components $\mathbf{F}$, named spectral profiles. They are used to extract the amplitude basis, hereafter referred to as amplitude envelopes according to 2.

$$\mathbf{E} = \mathbf{F} \cdot \mathbf{X} \tag{2}$$

The extracted amplitude envelopes present very salient beat detection functions with sharp peaks, sometimes accompanied by smaller peaks and plateaus stemming from crosstalk effects. The Percussiveness criterion [13] is computed from the amplitude envelopes in order to sort them according to their usefulness for beat spectrum computation. The accumulation of a detection function from ASE vectors is now replaced by a weighted sum of the amplitude amplitudes, whereas the corresponding Percussiveness serves as weighting factor. Further classification as described in [7], where symbolic drum patterns are derived, is circumvented here due to two reasons. On the one hand, a complete Transcription with detection and classification of the most salient percussive instruments would render the Beat Spectra useless, since it provides much more detailed high-level information. On the other hand, the final classification is error-prone and today not possible with convincing accuracy. Therefore the difficult step to the high-level domain is omitted and the intermediate results serve as basis for computation of the Beat Spectra.

Evaluation

In order to evaluate the benefit of the two preprocessing methods, tests were performed using the state-of-the-art rhythmic mid-level features described by Burred (BUR), Tzanetakis (TZA), Gouyon (GOU) as well as their combination (ALL). To derive a baseline score, they are extracted as described in the literature. Additionally, each of the two preprocessing methods is enabled to measure their influence.

Test Databases

An artificial and a real-world test-set of music pieces have been assembled. Both sets pose a genre classification task.

Artificial genre rhythm set: A number of different base rhythms were synthesized using the software sequencer Propellerhead's Reason [5]. The test set consists of 108 items from the following nine genres: Techno, R'n'B, Hard Rock, Glitch, Electro, Dub, Drum 'n Base, House and Hip-Hop. For each of these genres, two base rhythms were created and each was modified several times, from 90 BPM to 190 BPM in 20 BPM steps, obtaining 6 different tempi. Of course, solo drum patterns cannot be compared to complete music recordings, thus a second set was necessary.

Real-world genre set: The real-world test-set consists of 775 items from the following ten genres: Classical, Electronic, Jazz, Pop, Rock, German Pop, Urban, Speech, World and Miscellaneous. This material is of course much more diverse then the artificial set. Here, the preprocessing based on Drum Transcription is expected to have a positive influence on the results, since it attenuates instruments other than the percussive ones.

Evaluation Procedure

Gaussian Mixture Models (GMM) were used as classifiers. The classifiers did not use a rejection threshold. Applying conventional precision and recall measures to derive an overall recognition rate would lead to similar results for precision and recall. Thus, the pair-wise precision, recall and f-measure computation methods as proposed for measuring the accuracy of audio segmentation in [10] were applied. In case of the pair-wise precision and recall, the appearance of pairs of similar classes is evaluated, which indicates the overall results independently from the number of items in each class. The described test-sets were randomly split into 70% training and 30% test data. The splitting, GMM-training and classification was repeated 10 times in order to compute the mean precision (precis.), recall and f-measure as presented below.

Results and Discussions

Table 1 shows the average results achievable with the artificial (SYN) test-set, the original real-world test-set (RLW) and the Drum Transcription processed real-world test-set (RDT). The column Preproc. specifies if the logartihmic re-sampling was enabled (LOG) or not (RAW).

With the artificial test-set, a clear increase of all results can be observed when using the logarithmic re-sampling. The data from the artificial test set contained only drum loops. Therefore a comparison to the Drum Transcription preprocessing was not necessary, since the results can be considered as similar. The features proposed by Tzanetakis benefit from the tempo invariance given in the logarithmic domain. The results of Gouyon's and Burred's methods also improve which leads to the assumption, that spectral moments such as variance, kurtosis or skewness are per-se not tempo-invariant. They can compensate for stretching and translation of unimodal distributions, such as the Gaussian distribution. In contrast, Beat Spectra are strongly multimodal, since they represent the periodicities inherent in the music signal. The f-measure achievable with the combined

Data	Pre-proc.	Features	Precis.	Recall	F-measure
SYN	RAW	ALL	13.70	11.64	12.59
		BUR	21.48	19.80	20.60
		GOU	20.74	18.18	19.38
		TZA	9.26	7.72	8.42
	LOG	ALL	43.70	38.94	41.19
		BUR	42.59	36.98	39.59
		GOU	33.70	30.03	31.76
		TZA	36.67	33.45	34.98
RLW	RAW	ALL	26.2	25.3	25.8
		GOU	20.9	21.9	21.4
		BUR	22.1	24.8	23.4
		TZA	24.2	26.9	25.5
	LOG	ALL	23.7	25.8	24.7
		BUR	24.1	23.8	23.9
		GOU	21.3	18.5	19.8
		TZA	21.5	25.6	23.3
RDT	RAW	ALL	20.8	19.4	20.1
		BUR	18.1	20.6	19.2
		GOU	18.8	20.3	19.5
		TZA	19.2	21.5	20.3
	LOG	ALL	22.2	24.9	23.5
		BUR	22.9	24.5	23.6
		GOU	18.2	19.9	19.0
		TZA	20.0	22.2	21.0

Table 1: Evaluation results

feature vector containing all state-of-the-art measures showed the most promising improvement. However, the evaluation results with the real-world test set are less encouraging. The improvements achievable with the logarithmic re-sampling are in some cases only slightly better than the original methods, in the case that the Drum Transcription preprocessing has been applied. This preprocessing in turn decreases the overall results in comparison to the original signals. These results seemingly indicate that the proposed preprocessing methods are not beneficial for real-world music signals. The authors assume, that the reasons for these unsatisfactory results originate from the nature of real-world music. In contrast to the artificially synthesized drum-tracks, real music often shows rhythm changes. The raw Beat Spectra already exhibit a different shape if a single percussive instrument from the rhythm section stops playing for a certain period. In addition, there may be parts of songs that are not very important to the rhythm. They are neglected when modeling different genre classes with GMM and interpreting every available feature vector per song as representative. Furthermore, optical inspection of the raw Beat Spectra showed that there is much more variability inside the genre classes than simulated with the artificial test-set.

Conclusions

This publication compared state-of-the art rhythmic mid-level features with two preprocessing strategies intended to make these features more robust with regard to tempo changes and periodicity degradation. The results using an artificial music set show that the proposed method significantly increases the performance in a genre classification scenario. However, the results obtainable with a real-world genre classification data-set exhibited unsatisfactory results. In the future, test will be performed, in order to assess how the rhythmic information improves genre recognition in combination with other features. Additionally, tests will be conducted that take only a few feature vectors per song into account. Criteria such as the Percussiveness may additionally help to discard irrelevant instances for the genre classification.

References

[1] Burred, J; Lerch, A.: A Hierarchical Approach to Automatic Musical Genre Classification. In Proc. of DAFx-03 (2003)

[2] Dittmar, C.; Uhle, C.: Further Steps towards Drum Transcription of Polyphonic Music. In Proc. of 116th AES Convention (2004)

[3] Dittmar, C.; Bastuck C.; Gruhne, M.: Novel Mid-level Audio features for Music Similarity. In: Proc. of ICoMCS (2007)

[4] Foote, J; Cooper, M; Nam, U.: Audio Retrieval by Rhythmic Similarity. In: Proc. of ISMIR (2002)

[5] Propellerhead Reason 2.5, Future Music. Future Music **FM136** (2003), 30-37

[6] Gouyon, F. et al.: Evaluating rhythmic descriptors for musical genre classification. In Proc. of 25th AES International Conference (2004)

[7] Gruhne, M. et al.: Extraction of Drum Patterns and their Description within the MPEG-7 High-Level-Framework. In Proc. of ISMIR (2004)

[8] Gruhne, M. et al.: An Evaluation of Pre-Processing Algorithms for Rhythmic Pattern Analysis. In Proc. of 124th AES Convention (2008)

[9] Hyvärinen, A.; Karhunen, J.; Oja, E.: Independent Component Analysis, Wiley & Sons, New York, 2001

[10] Lukashevich, H.: Towards Quantitative Measures of Evaluating Song Segmentation. In Proc. ISMIR (2008)

[11] Plumbley, M.: Algorithms for Non-Negative Independent Component Analysis. IEEE Transactions on Neural Networks **14** (2003)

[12] Tzanetakis, G; Cook, P.: Musical Genre Classification of Audio Signals. In: IEEE transactions on Speech and Audio Processing **10** (2002), 293-302

[13] Uhle, C.; Dittmar, C.; Sporer, T.: Extraction of drum tracks from polyphonic music using independent subspace analysis. In Proc. of ICA (2003)

[14] Uhle, C. et al.: Low complexity musical meter estimation from polyphonic music. In Proc. of AES 25th International Conference (2004)