ON THE USE OF THE TEMPOGRAM TO DESCRIBE AUDIO CONTENT AND ITS APPLICATION TO MUSIC STRUCTURAL SEGMENTATION

Mi Tian, György Fazekas, Dawn A. A. Black, Mark Sandler

Centre for Digital Music, Queen Mary University of London
E1 4NS, London, UK

ABSTRACT

This paper presents a new set of audio features to describe music content based on tempo cues. Tempogram, a mid-level representation of tempo information, is constructed to characterize tempo variation and local pulse in the audio signal. We introduce a collection of novel tempogram-based features inspired by musicological hypotheses about the relation between music structure and its rhythmic components prominent at different metrical levels. The strength of these features is demonstrated in music structural segmentation, an important task in Music information retrieval (MIR), using several published popular music datasets. Results indicate that incorporating tempo information into audio segmentation is a promising new direction.

Index Terms— Audio signal processing, music segmentation, rhythm feature extraction, tempogram

1. INTRODUCTION

Automatic music structure analysis is one of the most important and difficult tasks in Music information retrieval (MIR) [1]. The main goal of music structural segmentation is to segment a piece into musically meaningful sections (e.g. verse, chorus). Characterizing music structure has its applications for instance in audio thumbnailing, music browsing and recommendation, as well as musicological research.

Techniques for music and audio segmentation mainly fall into three categories: the novelty-based, homogeneity-based, and the repetition-based approaches. The first one utilizes the hypothesis that segment boundaries can be characterized by prominent changes in audio features. These are typically detected using a "checkerboard" kernel correlated with the diagonal of the self-similarity matrix (SSM) of frame-wise audio features [2]. It can also be used as the first step in more complex segmentation methods. The homogeneity-based approach assumes stationarity in local statistical properties of features in structural segments. This may be modeled using machine learning and clustering techniques [3]. The principle of the repetition-based approaches is to find temporally ordered repetitions in the feature vectors or state sequences. Sections will then be characterized by some homogeneities detected in the investigated features or states [4]. These latter two approaches tend to have higher computational cost.

Various types of features have been proposed to compute the similarity capturing timbral, harmonic or rhythmic aspects of the input audio signal. Among these, rhythmic features are less frequently used [1, 5]. In this study, we will use novel rhythmic features derived from the tempo spectra of the audio signal, a two-dimensional representation indicating the strength of the local pulse over time. In related work, Grosche et.al. also point out the potential of integrating the concept of tempo representation into music structural segmentation [6]. Tempo-based features have also been used for cross-version novelty detection in [7].

The remainder of this paper is organized as follows. Section 2 introduces the methods and procedures for constructing the tempogram. Our new features are presented in Section 3. Section 4 deals with music structural segmentation using the presented features. Results and analyses are given in Section 5 and finally we conclude and provide some directions for future work in Section 6.

2. TEMPOGRAM

A tempogram is a time-pulse representation of an audio signal laid out such that it indicates the variation of pulse strength over time given a specific time lag l or a BPM value τ. The construction of a tempogram can be divided into two parts. The onset detection stage characterizes a series of musical events constituting the basic rhythmic content of the audio. This is followed by the estimation of local tempo using the autocorrelation or Fourier transform of the onset detection function (ODF) computed over short time windows [6, 8]. In this study, we are interested in variation in long term temporal structure, therefore we use the autocorrelation-based tempogram, because this emphasizes tempo subharmonics [6] corresponding to lower metrical levels.

2.1. Onset Detection

Similar to beat and tempo estimation systems, the tempogram relies on detecting sudden changes in the input signal corresponding to note onsets. Its quality therefore depends on the quality of the underlying ODF. Here we assume that the tempogram may be improved by using enhanced techniques for

audio onset detection. To this end, we adopt a method using the linearly weighted fusion of the complex domain (CD) and SuperFlux (SF) [9] onset detection functions. This method is proposed in a recent study, showing improved detection results over other onset detectors tested in a large-scale evaluation [10]. The calculation of the proposed ODF is given in Equation 1,

$$ODF(n) = \alpha \cdot CD(n) + (1 - \alpha) \cdot SF(n), \quad (1)$$

where n denotes the time index and α is the linear combination coefficient empirically set to 0.3 following the evaluation in [10].

2.2. Autocorrelation-based Tempogram

The calculation of the tempogram is based on the assumption that music exhibits coherent and locally periodic patterns. These patterns may be characterized by peaks in the autocorrelation function (ACF) of the ODF at certain time lags. To obtain an autocorrelation-based tempogram, we compute the local ACF of the ODF using a rectangular window W as shown in Equation 2.

$$A(t, l) = \frac{\sum_{n \in \mathbb{Z}} ODF(n) ODF(n + l) \cdot W(n - t)}{2N + 1 - l}, \quad (2)$$

for time $t \in \mathbb{Z}$ and time lag $l \in [0 : N]$ [6]. In the experiment, the time lag l corresponds to the tempo $\tau = 60/(r \cdot l)$ where $r = 0.005$ is the feature rate after resampling which improves time resolution for later processing steps. In our experiments, window sizes of 3, 5, 6 and 8s (with 0.2s overlap) were tested.

3. FEATURE DESCRIPTION OF AUDIO RHYTHMIC CONTENT

Rhythm information may enable the identification of structural elements in music that are not necessarily recognizable in the variation of timbre or harmony. Grosche et.al [11] introduced the Cyclic tempogram and the derived predominant local pulse (PLP) features. In this paper, we introduce additional methods to summarize the tempogram.

3.1. Dimensionality Reduction

Principal component analysis (PCA) is a multivariate data analysis technique that aims at minimizing the correlation between variables. It provides a linear orthogonal transformation into a new coordinate system such that after the projection the majority of variance lies in the first few dimensions and the variables become uncorrelated. In this work, the feature denoted TPCA is computed by using the first 20 principal components derived from the tempogram using PCA.

The Discrete cosine transform (DCT) can also be used as a dimensionality reduction technique. It is adopted as the last step in the calculation of the Mel-frequency cepstral coefficients (MFCCs) which proved highly successful in describing the timbral aspect of sound [12]. Inspired by this algorithm, we introduce a feature called *Tempogram cepstral coefficients* (TCC). For each tempogram frame we take the logarithm to emphasize the underline periodicity of the ACF, then calculate the DCT to obtain a compressed representation of the rhythmic content of the audio signal. The algorithm is illustrated in Equation 3.

$$TCC(n) = \sum_{l=0}^{N-1} log(A(l)) cos\left(\frac{\pi}{N}(l + \frac{1}{2})n\right), \quad (3)$$

where $n = 0, ..., N-1$. The DCT has the property to concentrate high energy components in the lower coefficients. Albeit the transform becomes orthogonal with appropriate scaling, we apply this to the log-compressed tempogram hence it enables a reduced representation focussing on the overall pulse regularity and helps to suppress noise. A 40 dimensional TCC is used in our experiments.

3.2. Band-wise Processing

In most music genres instruments play diverse roles in producing the overall rhythmic structure of a piece, thus each instrument can be prominent at different metrical level [13]. To capture this aspect we introduce two new aggregate feature types. *Tempo intensity* (TI) describes the strength of rhythmic components at different metrical levels. This feature is computed by aggregating the tempogram BPM bins into N bands using the following boundaries: $\{440, 240, 170, 130, 110, 90, 80, 65, 55, 40\}$ with $N = 9$, based on the assumption that tempo change is more salient over longer time periods. At time t when the tempo τ falls into the zth band ($z \in [0, N-1]$), we sum $A(t, \tau)$ to $T(z)$. Tempo components that drop out of this range will be discarded as they are presumed to have less contribution to the music structure.

Based on the assumption that small variations in less perceptually salient rhythmic components may help setting structural segments apart, we compress the band-wise intensity values using

$$TI(z) = T(z)^{\theta}, \quad (4)$$

where the exponent θ ($\theta \leq 0.5$) applies a fractional root function to $T(z)$. In the current work, we report results using the experimentally defined value of 0.4. This processing was inspired by and to some extent analogous to the calculation of perceived loudness in Moore's model [14].

The second feature, *Tempo intensity ratio* (TIR), describes the relative perceived salience of individual rhythmic components by calculating the intensity ratio of each band as defined above. The computation is given by Equation 5:

$$TIR(z) = \frac{T(z)}{\sum_{z=0}^{N-1} T(z)}. \quad (5)$$

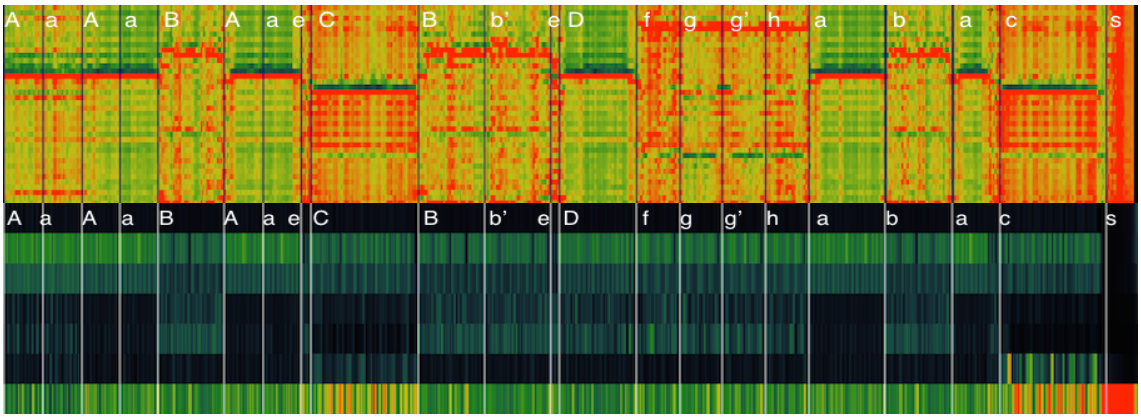

Fig. 1: Tempogram cepstral coefficients (TCC) (top) and Tempo intensity ratio (TIR) (bottom) for audio example Glamorama by Alex Q. The vertical lines and annotated texts indicate segment boundaries and section types.

In Fig. 1, two of the above introduced features of a piece of popular music are demonstrated in comparison with the annotated segment boundaries.

4. STRUCTURAL SEGMENTATION EXPERIMENT

The above described features (TCC, TPCA, TI, TIR) were assessed in the music structural segmentation task. First, tempogram features were extracted for two audio collections constituting our datasets. The features were then evaluated both individually and in combination. The rest of this section describes the datasets used for the evaluation as well as the segmentation process.

4.1. Dataset

In this study, we use two datasets with a large range of popular songs. The first is the SALAMI Internet archive dataset [15] denoted S-IA. It is composed of 272 popular pieces with a large variety of styles, regions and time periods. The second is the ISOPHONICS dataset from Queen Mary University of London[1], denoted ISO in the remainder of the paper. It is composed of 28 pieces of pop music, half of which are from The Beatles and the rest consists of Zweieck, Queen, Michael Jackson and Carole King. All audio files are sampled at 44.1 KHz and 16 bits per sample.

4.2. Similarity Analysis and Segmentation

A simple segment detection algorithm is used in this work inspired by [2]. A Self-similarity matrix (SSM) is constructed by calculating the pairwise Euclidean distance between vectors constituting the normalized feature matrix. A Gaussian-tapered checkerboard kernel is correlated along the main diagonal of the similarity matrix yielding a temporal novelty curve of the whole piece. In our experiments, a kernel size of 64 was used.

[1] `http://www.isophonics.net/datasets`

Post-processing and peak picking is then applied to the derived novelty curve to select prominent peaks as segment boundaries. To reduce noise that interferes with the selection of true boundaries, four processing techniques are applied: *i)* normalization using exponential weighting followed by, *ii)* zero-phase low-pass filtering for smoothing the novelty curve, *iii)* adaptive thresholding using a median filter and finally *iv)* polynomial fitting based peak selection, where coefficients of polynomials fitted around local maxima are used to accept or reject peaks. This post-processing and peak picking strategy follows our previous work [10].

5. RESULTS AND ANALYSIS

5.1. Segmentation Results

We use pairwise segment boundary recovery rate and median distance between the closest annotated boundaries and detected boundaries as indicators to evaluate the segmentation results. Detection rates under different tempogram window sizes are reported, as illustrated in Table 1. The S-IA dataset offers an informative comparison with the state of the art. Our results have surpassed the best results on the SALAMI dataset in the MIREX 2013 structural segmentation task (F=0.5193). We performed statistical significance test using the audio content in the overlapping set between S-IA and MIREX. Pairwise Wilcoxon Signed-Rank test shows significant improvement over the best performing algorithm [16] in MIREX[2] when the window length is set to 5 or 6s.

From Table 1 we can observe a notably higher recall (R) than precision rate (P) and this phenomenon is consistent under different sensitivity settings for boundary detection. This suggests that the tempogram features are relatively robust in

[2] MIREX structural segmentation 2013 results : `http://music-ir.org/mirex/wiki/2013:MIREX2013_Results` The Internet archive (S-IA) dataset and the MIREX testset are subsets of the SALAMI Dataset but they are not equivalent. Data for individual audio files were obtained from MIREX to provide a fair comparison in the statistical test.

	time window = **3**s					time window = **5**s					time window = **6**s					time window = **8**s				
	P	R	F	d_{AD}	d_{DA}	P	R	F	d_{AD}	d_{DA}	P	R	F	d_{AD}	d_{DA}	P	R	F	d_{AD}	d_{DA}
S-IA	0.5164***	0.4346***	0.4548	2.08	2.14	0.4483	0.8437***	0.5551***	0.73	2.18	0.4462	0.8542***	**0.5675*** **	0.70	2.16	0.5347***	0.5143***	0.5236	2.20	2.13
ISO	0.3828	0.3707	0.3716	2.55	2.17	0.3538	0.7949	**0.4757**	0.84	2.17	0.3532	0.7932	0.4756	0.77	2.10	0.4421	0.4813	0.4478	1.86	1.99

Table 1: Segmentation results on S-IA and ISO dataset. **P**, **R**, **F**: Segment boundary recovery precision, recall and F-measure rate measured at **3s**; $\mathbf{d_{AD}}$: Median distance from annotated segment boundaries to the closest detected boundaries; $\mathbf{d_{DA}}$: Median distance from detected segment boundaries to the closest annotated boundaries. *, ** and *** denote the presence of significant difference in P, R, and F between our results and the best performing algorithm in MIREX 2013 [16] for the SALAMI dataset at the level of 0.05, 0.01 and 0.001 using the Wilcoxon Signed-Rank test.

detecting true segment boundaries. However, a deficiency lies in the emergence of false positives, which leads to a lower precision hence bringing down the overall F-measure.

5.2. Discussion

The tempogram is built by using the local periodicity of the onset detection function, which identifies the amplitude, phase or other changes in the spectrogram of the input audio signal. Therefore, spectral information is not disregarded, rather, it is abstracted and reformed with the rhythmic cues emphasized. The fact that the features are tested on a collection of pop music instead of hand selected pieces with well defined rhythmic patterns indicates their general applicability to music content description.

To compare the performance of each feature used in the segmentation, we repeated the experiments by using each individual feature under the same experimental conditions on the S-IA dataset. The time window was set to 6s. Results are given by Table 2. All features exhibit quantitatively similar performances with an average F-measure of 0.5521 when the boundary recovery is measured at 3s. However, when measured at a finer scale (0.5s), a notable drop is observed. While the results obtained for all investigated features are better than the state of the art under a looser scale, they become worse at 0.5s. We can hypothesize this difference is due to the fact that the tempogram calculation utilizes larger window sizes compared to other algorithms reported in MIREX.

To investigate whether the difference in the performance of the four features is statistically significant, we used the Friedman test and obtained a p-value of 0.0008 for the F-measure at 3s and 0.3127 at 0.5s. This discrepancy may be explained by the grater variability of precision and recall under looser conditions, but does not allow for a consistent hypothesis about the complementary nature of the features.

The results indicate that the overall precision of the method introduced here could be improved. A balanced window length of 5 to 6s achieves better overall results and provides the best recall. However, we can observe from Table 1 that when the time window is set to 5s or 6s, lower precision is obtained compared to 3s or 8s, with the longest window yielding the highest precision. A possible explanation is that spurious peaks are suppressed due to longer window, though

	P(3s)	R(3s)	F(3s)	P(0.5s)	R(0.5s)	F(0.5s)	d_{AD}	d_{DA}
TCC	0.4450	0.8315	0.5488	0.1054	0.2021	0.1305	0.72	2.12
TPCA	0.4720	0.7759	0.5527	0.1065	0.1790	0.1253	0.82	2.17
TI	0.4511	0.8167	0.5525	0.1068	0.1986	**0.1313**	0.77	2.18
TIR	0.4658	0.7911	**0.5543**	0.1090	0.1913	0.1306	0.81	2.16

Table 2: Segmentation results on S-IA dataset using TCC, TPCA, TI and TIR (time window = 6s). **P(3s)**, **R(3s)**, **F(3s)**: Segment boundary recovery precision, recall and f-measure rate at 3s; **P(0.5s)**, **R(0.5s)**, **F(0.5s)**: Segment boundary recovery precision, recall and f-measure rate at 0.5s.

there is a drop in recall as less detail is exhibited in the SSM. The results may be improved by incorporating other musical information into the processing, for instance, by setting the window size in a tempo dependent fashion or using beat synchronous analysis windows.

6. CONCLUSION AND FUTURE WORK

In this paper, we studied a mid-level time-pulse representation of audio and presented a set of novel features to describe audio content. These features were applied to the music segmentation task and evaluated on two publicly available popular music databases. Our results indicate strong capacity of the tempogram features in describing music structure.

Aiming at a more comprehensive feature representation, in future work, we will combine the presented tempogram features with timbral, harmonic or other spectral features commonly used for audio segmentation. A validation of the selection of tempo bands and the hypothesis related to perceptual salience of long-term temporal structure will be carried out in future experiments. Promising directions lie also in the use of probabilistic models to generalize the feature space and suppress redundant information in the detection to improve precision. An investigation for systematic differences between the boundaries detected by our method and those in the annotations may also contribute to refining the estimation. Future work includes exploiting possible relationships between rhythm patterns using supervised learning. We also aim to investigate the applicability of tempogram to other research topics such as music genre recognition.

7. REFERENCES

[1] Jouni Paulus, Meinard Müller, and Anssi Klapuri, "State of the art report: Audio-based music structure analysis.," in *11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[2] Jonathan T Foote and Matthew L Cooper, "Media segmentation using self-similarity decomposition," in *Electronic Imaging*, 2003.

[3] Mark Levy and Mark B. Sandler, "Structural segmentation of musical audio by constrained clustering," in *Audio, Speech and Language Processing, IEEE Transaction on*, 2006.

[4] Masataka Goto, "A chorus-section detecting method for musical audio signals," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, 2003.

[5] Kristoffer Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," *EURASIP Journal on Applied Signal Processing*, 2007.

[6] Peter Grosche, M Müller, and Frank Kurth, "Cyclic tempogram—a mid-level tempo representation for musicsignals," in *Acoustics Speech and Signal Processing, IEEE International Conference on (ICASSP)*, 2010.

[7] Meinard Müller, Thomas Prätzlich, and Jonathan Driedger, "A cross-version approach for stabilizing tempo-based novelty detection.," in *13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[8] Matthew E.P. Davies and Mark D. Plumbley, "Causal tempo tracking of audio," in *5th International Conference on Music Information Retrieval (ISMIR)*, 2004.

[9] Sebastian Böck and Gerhard Widmer, "Maximum filter vibrato suppression for onset detection," in *16th International Conference on Digital Audio Effects (DAFx). Maynooth, Ireland*, 2013.

[10] Mi Tian, György Fazekas, D. A. A. Black, and Mark Sandler, "Design and evaluation of onset detectors using different fusion policies," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014.

[11] Peter Grosche and Meinard Müller, "Extracting predominant local pulse information from music recordings," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2011.

[12] Steven Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1980.

[13] Richard Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music Perception*, 1994.

[14] Brian C.J. Moore, Brian R. Glasberg, and Thomas Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, 1997.

[15] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie, "Design and creation of a large-scale database of structural annotations.," in *12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[16] Bruno Rocha, Niels Bogaards, and Aline Honingh, "Detection of structural boundaries in electronic dance music," in *Mirex 2013*.