ACOUSTIC FEATURES FOR MUSIC PIECE STRUCTURE ANALYSIS

Jouni Paulus, Anssi Klapuri[*]

Department of Signal Processing
Tampere University of Technology
Tampere, Finland
{jouni.paulus,anssi.klapuri}@tut.fi

ABSTRACT

Automatic analysis of the structure of a music piece aims to recover its sectional form: segmentation to musical parts, such as chorus or verse, and detecting repeated occurrences. A music signal is here described with features that are assumed to deliver information about its structure: mel-frequency cepstral coefficients, chroma, and rhythmogram. The features can be focused on different time scales of the signal. Two distance measures are presented for comparing musical sections: "stripes" for detecting repeated feature sequences, and "blocks" for detecting homogenous sections. The features and their time scales are evaluated in a system-independent manner. Based on the obtained information, the features and distance measures are evaluated in an automatic structure analysis system with a large music database with manually annotated structures. The evaluations show that in a realistic situation, feature combinations perform better than individual features.

1. INTRODUCTION

Musical pieces typically employ a lot of repetition and variation at different time scales: from the re-occurrences of melodic phrases to the repetitions of musical parts of tens of seconds in length. This holds the piece together and imposes a certain structure on it. Here, the analysis of the structure of a musical piece means recovering a description of the sectional form of the piece. This means recovering a temporal segmentation to parts like intro, chorus and verse, and grouping segments that are occurrences of the same musical part. Segments that are occurrences of the same musical part are said to belong to the same *group*.

Knowledge of the structure of a music piece enables several applications. Perhaps the most obvious is a content-aware music player which allows the user to navigate within the piece based on the structural information [3]. Others include remixing the piece, creating a mash-up from several pieces, or utilising the acoustic similarity of musical part occurrences in audio coding.

Different aspects related to structural analysis have been discussed in the literature. Locating one or more of the occurrences of the chorus section is a popular aim because usually the chorus is a relatively good thumbnail or a representative sample of the piece [3, 2, 8]. Some systems only aim to detect boundaries of musical parts [10], and some others additionally perform the segment grouping [6].

There are some commonalities in the features used by the systems: the timbral content is often described using mel-frequency cepstral coefficients (MFCCs) and the harmonic content with chroma. Some existing systems assume that the features (acoustic characteristics) of a piece stay somewhat constant during each occurrence of a musical part and change at part boundaries. Some other systems assume that during a part occurrence, the features form a sequence which then re-occurs when the part is repeated.

This paper considers the task of describing the whole structure of a piece including segmentation to musical parts and grouping the different occurrences of a part. In addition to MFCCs and chroma, a third feature, rhythmogram, describing the rhythmic content is considered. The features are compared in the task of grouping sections that are occurrences of the same part. The statistics of the features are analysed in order to obtain an algorithm-independent view of what features are the most informative and what time scale should be employed. The features and distance measures are compared using a large data set of popular music pieces with manually annotated structures.

The usefulness of each individual feature and their combinations is evaluated in an automatic structure analysis system. The analysis results allow us to determine the features and the distance measures that contain useful information.

2. FEATURES AND DISTANCE MEASURES

As with all pattern recognition tasks, the used features are a key factor in the overall system performance. A study of the perceptual cues involved in the perception of structural boundaries suggests that there are several factors involved, e.g., repetitions, and changes in timbre and rhythm [1]. The features and distance measures that are evaluated aim to address these factors: MFCCs for timbre, harmonic content with chroma, and rhythmic changes should become noted with the use of rhythmogram. A similar set of features has been used earlier in [4], but the work evaluated the performance of individual features with one distance measure.

2.1. Acoustic Features

The feature extraction starts by estimating the locations of musical beats. This is done for anchoring the features to the time grid of the music. Also, the beat-synchronised frames make the resulting features tempo invariant as the frame length is adjusted along with tempo changes. The analysis is done using the method described in [5], with some modifications to improve the tempo stability. After the analysis, the period of the estimated beat is halved by inserting extra beats between each two found beats in order to remove the effect of possible π-phase errors in the estimated pulse.

The general timbral characteristics of the signal are described with MFCCs. They are calculated in 92.9 ms frames with 46.4 ms overlap. A vector of sub-band energies at 42 sub-bands is discrete

[*] This work was supported by the Academy of Finland, project No. 5213462 (Finnish Centre of Excellence program 2006 - 2011).

cosine transformed, the lowest coefficient is discarded and the following 12 are used as the feature vector for the frame.

Harmonic content of the signal is described with chroma. It is calculated using the method from [9] and the same frame blocking parameters as in the MFCC calculation. First, the saliences of different fundamental frequencies in the range 80–640 Hz are estimated. Then the frequency scale is transformed into a musical one by retaining only the maximum-salience fundamental frequency component within each semitone bin. Finally, saliences of the octave equivalence classes are summed to produce a 12-dimensional chroma vector.

The rhythmic content of the music is described with rhythmogram [4]. The calculation method is modified by replacing the perceptual spectral flux front-end with musical accent estimation front-end from the meter analysis system [5]. The purpose of the front-end is to produce a signal which reacts to onsets in the input signal. The used accent signal is a by-product of the meter analysis process and it has proved to be effective.

The rhythmogram calculation procedure is the following. First, the mean of the accent signal is removed, then autocorrelation of the signal is calculated in long (several seconds) overlapping windows, and the autocorrelation values between 0 and 2 seconds are retained. The values are normalised to produce value 1 at lag 0. The length of the autocorrelation window is determined by the time scale the feature is focused on, as will be described later.

2.2. Temporal Filtering and Self-distance Matrices

All the calculated features are transformed to beat-synchronised time grid by calculating the mean value of the feature in each beat frame. The resulting feature vector for the beat frame k is denoted by $\mathbf{f_k}$. The same notation is used for all the features.

The use of multiple time scales has been noted to be beneficial for structure analysis [8, 10]. The time scale of the MFCCs and chroma features is varied by low-pass filtering the features along time. The filter is a 2nd order Butterworth IIR filter with cut-off frequency that determines the time scale. The filter is applied to the feature time-series twice: forward and backward in time, in order to double the filter's magnitude response and to cancel phase distortions. The filter cut-off frequency is $\omega_c = 1/\tau$, where τ is the time scale parameter. The filtering is not done for the rhythmogram feature. Instead, the length of the autocorrelation window N is varied depending on the inspected time scale. After the filtering, the features are normalised to have zero mean and unity variance over the piece.

A *self-distance matrix* (SDM) is used to detect repeated sections. An SDM is a square matrix with as many rows and columns as there are frames in the signal (here, beat-synchronised frames). Each element $D_{k,l} = d\left(\mathbf{f_k}, \mathbf{f_l}\right)$ in the matrix denotes the distance of the corresponding frames k and l in the signal calculated using the cosine distance measure.

Figure 1 illustrates six SDMs calculated for the piece "Tuonelan koivut" by Kotiteollisuus.[1] The darker stripes directed -45 degrees in the right column SDMs indicate sequences that are repeated during the piece. There are blocks of low distance in left column SDMs approximately at the same locations as there are stripes in the right-hand-side SDMs. They indicate features remaining approximately constant during the section and its repeated occurrences. These stripes and blocks motivate the used distance measures between segments.

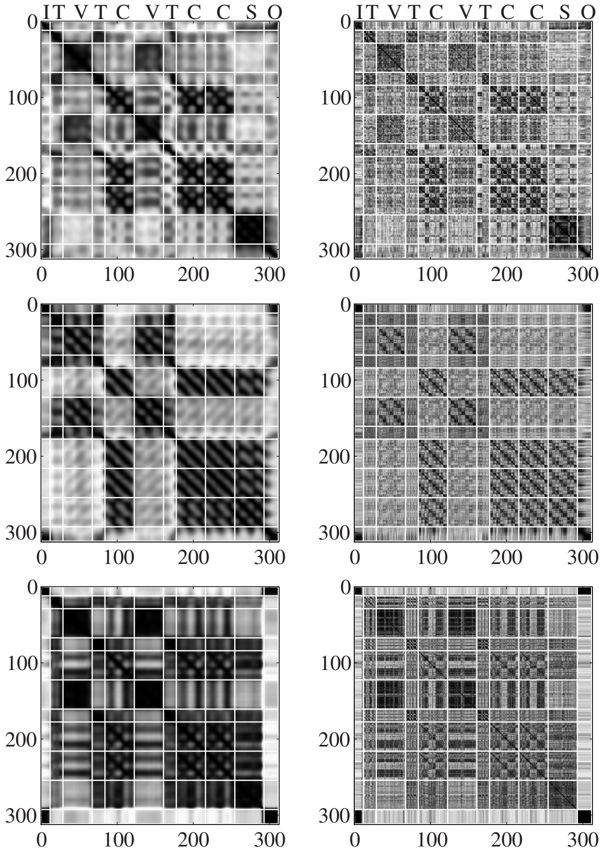

Figure 1: *Examples of SDMs for (top to bottom) MFCCs, chroma and rhythmogram, illustrating the formed blocks and stripes. The matrices on the left are calculated using lower low-pass cut-off frequency, while the matrices on the right are calculated using higher cut-off frequency. All axes show time in seconds, and darker pixel value denotes lower distance. The overlaid white grid illustrates the annotated part borders, and the annotated part labels are indicated above the top panel: intro (I), theme (T), verse (V), chorus (C), solo (S), outro (O).*

2.3. Distance Measures

Two distance measures for comparing segments s_m and s_n are formulated. The frames of the segments define a submatrix $\tilde{D}_{m,n}$ of the SDM. *Block distance* measure $d_B\left(s_m, s_n\right)$ is defined as the average element value in $\tilde{D}_{m,n}$. *Stripe distance* measure $d_S\left(s_m, s_n\right)$ is defined by calculating the path of the lowest cumulative distance across the submatrix $\tilde{D}_{m,n}$. The used local path constraint forces the path to make one step in one or both dimensions.

3. SYSTEM-INDEPENDENT ANALYSIS OF FEATURES AND TIME SCALES

The usefulness of the features and defined distance measures was evaluated independently of the structure analysis system.

[1] Note that the piece was selected for illustration because the matrices show very prominent stripes and blocks; this is not often the case.

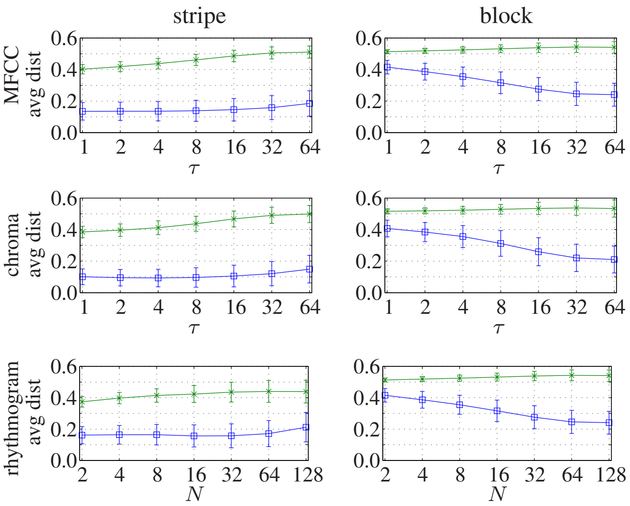

Figure 2: *Average stripe and block distance values for segment pairs calculated from different features with varying low-pass cut-off parameter τ (for MFCCs and chroma) or autocorrelation window length N in beats (rhythmogram). Each panel illustrates the average distances between segments from the same group (□), and from different groups (×). The error bars illustrate the standard deviation around the mean.*

3.1. Data

The proposed distance measures and temporal processing are evaluated using a large data set of popular music pieces with structural annotations. The dataset consists of 557 pieces from Western popular music genre, mainly of pop/rock, but also more diverse data such as jazz and blues are present.[2] The pieces were selected to provide a representative sample of the music that is being played on the radio. The annotation of a piece contains temporal segmentation to musical parts and labelling of the segments using the name of the part. The annotations were done by two research assistants with some musical background.

3.2. Results

Unlike earlier experiments, where the block-like properties have been searched from SDMs from MFCCs, and stripes have been searched from SDMs from chroma, both distance measures are applied on all three features. The parameter affecting the time scale is varied (the low-pass cut-off with MFCCs and chroma, autocorrelation window length with rhythmogram). The goodness of a feature and distance measure pair was evaluated by calculating the average distance between segments from the same group and segments from different groups in the data set. The knowledge about the segment locations and their groupings was taken from the annotations.

The effect of the time scale parameter on the features is illustrated in Figure 2. The graphs show the overall average values for the distance measures d_S and d_B calculated over the whole data set. The larger the gap between inter- and intra-group distances is, the better the temporal parameter value is for the distance measure.

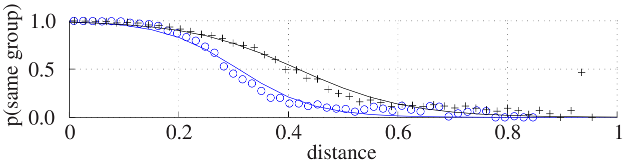

Figure 3: *Sigmoidal mapping function from distances between segments to probability that the two segments are from the same group. The solid lines are the estimated sigmoids. The empirical probabilities for the stripe distance are denoted with ○, and for the block distances with +. The given sigmoids are from MFCCs, but the ones for chroma and rhythmogram are very similar.*

The final time scale choice for each of the features and distance measures was done by approximating the distributions of average inter- and intra-group distances as Gaussians and selecting the time scale value that minimised the overlap between the two distributions. This was done for each of the features separately. The selected values for MFCC τ are 8 for d_S and 64 for d_B, for chroma the values are 0.5 and 64, and the rhythmogram autocorrelation window length corresponding 4 and 32 beat frames.

4. COMPARISON OF FEATURES AND DISTANCE MEASURES

The practical usability of the calculated distances for segment pairs are evaluated by using them in a structure analysis system described in more detail in [7]. The distances are transformed to probabilities of the two segments to be occurrences of the same musical part $p(s_m, s_n)$. This is done by fitting a sigmoidal function to the calculated distance values with logistic regression. Two of the resulting sigmoids are illustrated in Figure 3. The probability values based on different features and distance measures are combined by calculating a geometric mean of the probabilities involved.

The analysis system creates different structural descriptions E covering the whole piece, and evaluates the fitness of the descriptions using the segment pair probabilities in:

$$P(E) = \sum_{m=1}^{M}\sum_{n=1}^{M} A(s_m, s_n) L(s_m, s_n), \tag{1}$$

where

$$L(s_m, s_n) = \begin{cases} \log\left(p(s_m, s_n)\right) & \text{if } g_m = g_n \\ \log\left(1 - p(s_m, s_n)\right) & \text{if } g_m \neq g_n \end{cases}. \tag{2}$$

The weighting factor $A(s_m, s_n)$ corresponds to the number of elements in the submatrix $\tilde{D}_{s_m,s_n}$. It is motivated by the need to cover the whole SDM and to enable comparing descriptions with different number of segments. g_m defines the group to which the segment s_m has been assigned to, and M is the number of segments in the description. The description with the largest value of $P(E)$ is the analysis result. The search algorithm is omitted here due space restrictions.

Two evaluation schemes are used. First, the segment boundaries are taken from the annotations, and the system has to find the grouping. In the second, different segmentations are generated from the acoustic data using the novelty measure [2]. The novelty measure is calculated from all features separately and then summed.

[2] Full list of pieces is available at <http://www.cs.tut.fi/sgn/arg/paulus/TUTstructure07_files.html>.

Different feature and distance measure combinations are tested using 10-fold cross-validation scheme. On each iteration, 90% of the data is used to determine the warping function parameters, while the remaining 10% is used for testing. The presented results are averaged over all folds.

4.1. Evaluation Measure

The used measure has earlier been used in evaluation of structure analysis system in [6]. It considers all pairs of frames and whether or not the frames are assigned in the same group. The metrics are calculated from the pairs of frames that are assigned to the same group. Recall rate R_r is the ratio of correct assignments to assignments in the ground truth, and precision rate R_p is the ratio of correct assignment to made assignments. F-measure is calculated from these two as $F = 2R_pR_r/(R_p + R_r)$.

4.2. Results

The evaluation of individual feature/distance pairs and of the best combinations of 2–6 pairs are given in Table 1. The six leftmost columns indicate the used features: MFCCs (M), chroma (C), rhythmogram (R), and distance measure: block (B), stripe (S). A dot in the cell indicates the feature/distance was used. The three rightmost columns give F-measure, precision and recall rates. A performance increase of 1 percentage unit in the F-measure can be considered to be statistically significant (p < 5%).

In the grouping task, stripe distance measure performs well: MFCC and chroma stripes alone are better than any combination of 4–6 features/distances. This is probably because when the segment boundaries are given, the repetitions form distinct diagonal stripes across the submatrices. Considering the combinations, chroma stripe, rhythmogram block and MFCC block/stripe are present, confirming to the observations in [1]. Without a given segmentation the stripe performance decreases: it still has better precision than blocks, but worse recall rate. This means that the recurrence stripes are not found so often. Also, combinations improve the result over individual results. In view of the F-measure, the performance of the best feature combination is similar to the results presented in [6], although it should be noted that the data sets differ.

Table 1: *Results for different features and distance measures in both evaluation schemes. Values in the top half are for given segment boundaries and values in the bottom half are for a fully automatic system. See text for other details.*

MB	MS	CB	CS	RB	RS	F(%)	R_p(%)	R_r(%)
●						80.9	84.5	80.9
	●					86.1	93.7	82.1
		●				77.4	78.5	80.7
			●			85.8	91.6	83.7
				●		73.9	75.1	78.8
					●	78.8	90.0	74.2
	●		●			86.2	95.6	81.1
●			●	●		85.1	91.7	82.2
●	●		●	●		85.5	93.7	81.2
●	●	●	●	●		84.9	92.7	81.0
●	●	●	●	●	●	84.4	93.4	79.8
●						58.5	54.1	68.2
	●					56.0	65.8	52.8
		●				53.5	46.5	69.3
			●			53.1	67.1	48.6
				●		51.0	45.0	68.5
					●	47.4	61.4	45.5
●	●					60.4	60.2	65.2
●		●	●			60.4	61.2	64.5
●		●	●		●	59.9	64.7	60.8
●	●	●	●	●		61.7	65.9	63.1
●	●	●	●	●	●	60.3	67.7	59.7

5. CONCLUSIONS

Three acoustic features: MFCCs, chroma and rhythmogram and two distance measures were evaluated in the task of detecting repeated sections in a music piece. Moreover, the effect of focusing the features on different time scales was evaluated. Finally, a fully automatic music structure analysis system was presented and evaluated on a large database of manually annotated popular music pieces. The results suggest that with reliable segmentation "stripe" distance measure with MFCCs or chroma work well alone, with unreliable segmentation, combinations of features and distance measures should be employed.

6. REFERENCES

[1] M. J. Bruderer, M. McKinney, and A. Kohlrausch. Structural boundary perception in popular music. In ISMIR, Victoria, Canada, 2006.

[2] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In WASPAA, New Platz, New York, USA, 2003.

[3] M. Goto. A chorus-section detecting method for musical audio signals. In ICASSP, Hong Kong, 2003.

[4] K. Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Advances in Signal Processing*, 2007.

[5] A. Klapuri, A. Eronen, and J. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[6] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008.

[7] J. Paulus and A. Klapuri. Music structure analysis with probabilistically motivated cost function with integrated musicological model. In ISMIR, Philadelphia, Pennsylvania, USA, 2008.

[8] G. Peeters. Deriving musical structure from signal analysis for music audio summary generation: "sequence" and "state" approach. In *Lecture Notes in Computer Science*, vol. 2771. Springer-Verlag, 2004.

[9] M. P. Ryynänen and A. P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 2008.

[10] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In ISMIR, Vienna, Austria, 2007.