MUSIC STRUCTURE ANALYSIS BY RIDGE REGRESSION OF BEAT-SYNCHRONOUS AUDIO FEATURES

Yannis Panagakis and Constantine Kotropoulos
Department of Informatics
Aristotle University of Thessaloniki
Box 451 Thessaloniki, GR-54124, Greece
{panagakis,costas}@aiia.csd.auth.gr

ABSTRACT

A novel unsupervised method for automatic music structure analysis is proposed. Three types of audio features, namely the mel-frequency cepstral coefficients, the chroma features, and the auditory temporal modulations are employed in order to form beat-synchronous feature sequences modeling the audio signal. Assume that the feature vectors from each segment lie in a subspace and the song as a whole occupies the union of several subspaces. Then any feature vector can be represented as a linear combination of the feature vectors stemming from the same subspace. The coefficients of such a linear combination are found by solving an appropriate ridge regression problem, resulting to the ridge representation (RR) of the audio features. The RR yields an affinity matrix with nonzero within-subspace affinities and zero between-subspace ones, revealing the structure of the music recording. The segmentation of the feature sequence into music segments is found by applying the normalized cuts algorithm to the RR-based affinity matrix. In the same context, the combination of multiple audio features is investigated as well. The proposed method is referred to as ridge regression-based music structure analysis (RRMSA). State-of-the-art performance is reported for the RRMSA by conducting experiments on the manually annotated Beatles benchmark dataset.

1. INTRODUCTION

The structural description of a music piece at the time scale of segments, such as intro, verse, chorus, bridge, etc. is referred to as the *musical form* of the piece [15]. Its derivation from the audio signal is a core task in music thumbnailing and summarization, chord transcription [10], learning of music semantics and music annotation [1], song segment retrieval [1], or remixing [6].

Human listeners analyze and segment music into meaningful parts by detecting the structural boundaries between the segments thanks to the perceived changes in timbre, tonality, and rhythm over the music piece. Music structure analysis extracts low-level feature sequences from the audio signal in order to model the timbral, melodic, and rhythmic content [15]. The segmentation of the feature sequences into structural parts is performed by employing methods based on either repetition, homogeneity, or novelty [1, 6, 7, 12, 14, 15, 17] to analyze a recurrence plot or a self-similarity distance matrix. For a comprehensive review on automatic music structure analysis systems the interested reader is referred to [4, 15] (and the references therein).

In this paper, a novel method for music structure analysis is proposed, which differs significantly from the previous methods. In particular, three types of audio features, namely the *mel-frequency cepstral coefficients* (MFCCs), the *Chroma* features, and the *auditory temporal modulations* (ATMs) are employed in order to form beat-synchronous feature sequences modeling the timbral, tonal, and rhythmic content of the music signal. It is reasonable to assume that due to the timbral, tonal, and rhythmic homogeneity within the music segments, the audio features extracted from a specific music segment are highly correlated and thus linearly dependent. Therefore, there is a linear subspace that spans the beat-synchronous audio features for each specific music segment implying that the sequence of feature vectors extracted from the whole music recording will lie in a union of as many linear subspaces as the music segments of this recording are. Accordingly, a feature vector extracted at the time scale of a beat can be represented as a linear combination of the feature vectors stemming from the subspace it belongs to. Formally, one solves an appropriate inverse problem in order to obtain the representation of each feature vector with respect to a dictionary, which is constructed by all the other feature vectors as atoms (i.e., column vectors). Here, it is proved that the joint *ridge representation* (RR) of the features drawn from a union of independent linear subspaces exhibits nonzero within-subspace affinities and zero between-subspace affinities. That is, a *ridge regression* [1] problem is solved, which admits a unique and closed-form solution. The segmentation of the feature sequence into music segments is revealed by applying the normalized cuts spectral clustering algorithm [16] to the RR-based affin-

[1] Ridge regression is also known as Tikhonov regularization.

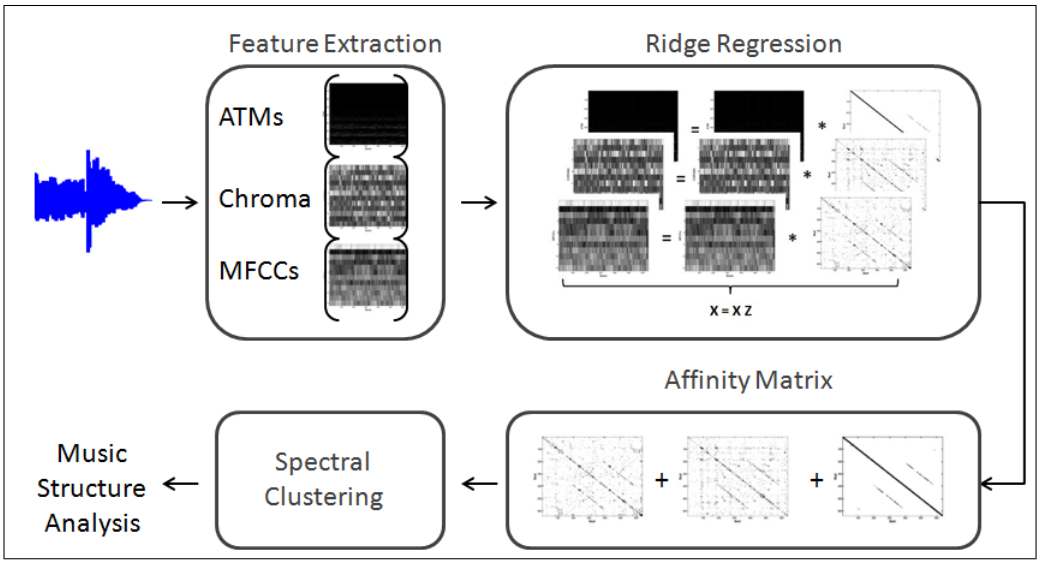

Figure 1. Each music recording is modeled by three audio features, namely the MFCCs, the Chroma features, and the ATMs resulting to three beat-synchronous feature matrices. The RR is derived for each feature matrix and three affinity matrices are obtained as described in Section 3. A cross-feature affinity matrix is obtained by linearly combining the affinity matrices obtained for the individual features. The segmentation of the music recording into music segments is obtained by applying the normalized cuts spectral clustering algorithm to the cross-feature RR-based affinity matrix.

ity matrix. Provided that music segments can seldom be revealed efficiently by resorting to a single feature, multiple features are extracted from each music recording and the cross-feature information is utilized in order to obtain a reliable music segmentation. To this end, a cross-feature RR-based affinity matrix is constructed by linearly combining the RR-based affinity matrices obtained for each individual feature. Again, the segmentation of the feature sequence into music segments is obtained by applying the normalized cuts to the cross-feature RR-based affinity matrix. The proposed method is referred to as *ridge regression-based music structure analysis* (RRMSA) and it is outlined in Fig. 1.

The performance of the RRMSA is assessed by conducting experiments on the manually annotated Beatles dataset. The RRMSA is demonstrated to yield a state-of-the-art performance.

The remainder of the paper is as follows. In Section 2, the audio features employed are briefly described. The RRMSA is detailed in Section 3. Dataset, evaluation metrics, and experimental results are presented in Section 4. Conclusions are drawn in Section 5.

2. AUDIO FEATURE REPRESENTATION

The variations between different music segments are captured by extracting three audio features from each monaural music recording sampled at 22.05-kHz. In particular, the MFCCs, the Chroma features, and the ATMs are employed.

1) The MFCCs encode the timbral properties of the music signal by parameterizing the rough shape of spectral envelope. Following [14], the MFCC calculation employs frames of duration 92.9 ms with a hop size of 46.45 ms and a 42-band filter bank. The correlation between the frequency bands is reduced by applying the discrete cosine transform along the log-energies of the bands. The zeroth order coefficient is discarded yielding a sequence of 12-dimensional MFCCs vectors.

2) The Chroma features are able to characterize the harmonic content of the music signal by projecting the entire spectrum onto 12 bins representing the 12 distinct semitones (or chroma) of a musical octave. They are calculated by employing 92.9 ms frames with a hope size of 23.22 ms as follows. First, the salience of different fundamental frequencies in the range $80 - 640$ Hz is calculated. The linear frequency scale is transformed into a musical one by selecting the maximum salience value in each frequency range corresponding to one semitone. Finally, the octave equivalence classes are summed over the whole pitch range to yield a sequence of 12-dimensional chroma vectors.

3) The ATMs carry important time-varying information of the audio signal [11]. They are obtained by modeling the path of human auditory processing as a two-stage process. In the first stage, which models the early auditory system, the acoustic signal is converted into a time-frequency distribution along a logarithmic frequency axis, the so-called *auditory spectrogram*. The early auditory system is modeled by Lyons' passive ear model [9] employing 96 frequency channels ranging from 62 Hz to 11 kHz. The auditory spectrogram is then downsampled along the time axis in order to obtain 10 feature vectors between two successive beats. The underlying temporal modulations of the music signal are derived by applying a biorthogonal wavelet filter along each temporal row of the auditory spectrogram, where its mean has been previously subtracted, for 8 discrete rates $r \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ Hz ranging from slow to fast temporal rates. Thus, the entire

auditory spectrogram is modeled by a three-dimensional representation of frequency, rate, and time which is then unfolded [2] along the time-mode in order to obtain a sequence of $96 \times 8 = 728$-dimensional ATMs.

Postprocessing. Sequences of *beat-synchronous* feature vectors are obtained by averaging the feature vectors over the beat frames. The latter are found by using the beat tracking algorithm described in [5]. Each row of the beat-synchronous feature matrix is filtered by applying an average filter of length 8. Finally, each feature vector undergoes a normalization in order to have zero-mean and unit ℓ_2 norm.

3. MUSIC STRUCTURE ANALYSIS BASED ON RIDGE REGRESSION

In this section, the RRMSA is detailed. Let a given music recording of K music segments be represented by a sequence of N beat-synchronous audio feature vectors of size d, i.e., $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\dots|\mathbf{x}_N] \in \mathbb{R}^{d\times N}$. The perceived timbral, tonal, and rhythmic homogeneity within a music segment implies that the audio features extracted from this music segment are highly correlated, exhibiting linear dependence. This motivated us to assume that beat-synchronous feature vectors belonging to the same music segment live into the same subspace. Therefore, if a music recording consists of K music segments, the sequence of N beat-synchronous audio feature vectors (i.e., the columns of $\mathbf{X}$) are drawn from a union of K independent linear subspaces of unknown dimensions. Thus, each feature vector can be represented as a linear combination of feature vectors drawn from the same subspace. That is, $\mathbf{X} = \mathbf{XZ}$, where $\mathbf{Z} \in \mathbb{R}^{N\times N}$ is the representation matrix, which contains the linear combination coefficients in its columns [3]. Clearly, $z_{ij} = 0$, if $\mathbf{x}_i$ and $\mathbf{x}_j$ lie on different subspaces and nonzero otherwise.

Such a representation matrix $\mathbf{Z}$ can be found by solving a least-squares problem regularized by the Frobenius norm (denoted by $\|.\|_F$), the so-called *ridge regression* problem:

$$\underset{\mathbf{Z}}{\operatorname{argmin}} \quad \|\mathbf{X} - \mathbf{X}\,\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_F^2. \tag{1}$$

The unique solution of the unconstrained convex problem (1) is referred to as *ridge representation* (RR) matrix and it is given in closed-form by:

$$\mathbf{Z} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}. \tag{2}$$

Technically, the desired property of the RR matrix to admit nonzero entries for within-subspace affinities and zero entries for between-subspace affinities is enforced by the regularization term $\lambda\|\mathbf{Z}\|_F^2$ in (1) as proved in Theorem 1, which is a consequence of Lemma 1. This result indicates that if the data follow subspace structures (i.e., come from a union of independent subspaces), the correct identification of the subspaces can be obtained accurately, fast, and in closed form by solving (1) without imposing sparsity or other constraints on the data model.

Lemma 1 [2]. For any four matrices $\mathbf{B}, \mathbf{C}, \mathbf{D}$, and $\mathbf{F}$ of compatible dimensions,

$$\left\|\begin{bmatrix}\mathbf{B} & \mathbf{C}\\ \mathbf{D} & \mathbf{F}\end{bmatrix}\right\|_F^2 \geq \left\|\begin{bmatrix}\mathbf{B} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}\end{bmatrix}\right\|_F^2 = \|\mathbf{B}\|_F^2 + \|\mathbf{F}\|_F^2. \tag{3}$$

Theorem 1. Assume the columns of $\mathbf{X}$ (i.e., the feature vectors) are drawn from a union of K linear independent subspaces of unknown dimensions. Without loss of generality, one may decompose $\mathbf{X}$ as $[\mathbf{X}_1|\mathbf{X}_2|\dots|\mathbf{X}_K] \in \mathbb{R}^{d\times N}$, where the columns of $\mathbf{X}_k \in \mathbb{R}^{d\times N_k}$, $k = 1, 2, \dots, K$ correspond to the N_k feature vectors originating from the kth subspace. The minimizer of (1) is block-diagonal.
The proof can be found in the Appendix.

Let $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the singular value decomposition (SVD) of $\mathbf{Z}$. Set $\tilde{\mathbf{U}} = \mathbf{U}(\mathbf{\Sigma})^{\frac{1}{2}}$ and $\mathbf{M} = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T$. A RR-based nonnegative symmetric affinity matrix $\mathbf{W} \in \mathbb{R}_+^{N\times N}$ has elements [8]:

$$w_{ij} = m_{ij}^2. \tag{4}$$

The segmentation of the columns of $\mathbf{X}$ into K clusters (i.e., music segments) is performed by employing the normalized cuts [16] onto the RR-based affinity matrix $\mathbf{W}$.

Since the music segments cannot be accurately derived by resorting to one feature, cross-feature information is expected to produce a more reliable music segmentation. Let $\mathbf{W}_m$, $\mathbf{W}_c$, and $\mathbf{W}_a$ be the RR-based affinity matrix obtained, when the MFCCs, the Chroma, and the ATMs are employed, respectively. A cross-feature RR-based affinity matrix $\mathbf{W}_{cf} \in \mathbb{R}_+^{N\times N}$ can be constructed by:

$$\mathbf{W}_{cf} = 1/3(\mathbf{W}_m + \mathbf{W}_c + \mathbf{W}_a), \tag{5}$$

or any other combination of the individual affinity matrices. The segmentation of the music recording can be obtained by applying the normalized cuts [16] to the cross-feature RR-based affinity matrix $\mathbf{W}_{cf}$.

In general, the number of segments K in a music recording is unknown and thus it is reasonable to be estimated. To this end, the soft-thresholding approach is employed [8]. That is, the estimated number of segments $\bar{K}$ is found by:

$$\bar{K} = N - \text{int}(\sum_{i=1}^{N} f_\tau(\sigma_i)). \tag{6}$$

The function int(.) returns the nearest integer of a real number, $\{\sigma_i\}_{i=1}^N$ denotes the set of the singular values of the Laplacian matrix derived by the corresponding affinity matrix, and f_τ is the soft-thresholding operator defined as $f_\tau(\sigma) = 1$ if $\sigma \geq \tau$ and $\log_2(1 + \frac{\sigma^2}{\tau^2})$, otherwise. Clearly, the threshold $\tau \in (0, 1)$.

[2] The tensor unfolding can be implemented in Matlab by employing the `tenmat` function of the MATLAB Tensor Toolbox available at: `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`.

[3] Due to the assumptions stated at the beginning of Section 3, the matrix $\mathbf{X}$ does not has full column rank. Therefore, $\mathbf{X} = \mathbf{XZ}$ does not admit the identity matrix as solution.

4. EXPERIMENTAL EVALUATION

4.1 Dataset, Evaluation Procedure, and Evaluation Metrics

Beatles dataset[4]: The dataset consists of 180 songs by The Beatles. The songs are annotated by the musicologist Alan W. Pollack. Segmentation time stamps were inserted at Universitat Pompeu Fabra (UPF). Each music recording contains on average 10 segments from 5 unique classes [17].

The structure segmentation is obtained by applying the RRMSA to each individual feature sequence as well as to all possible feature combinations. In Fig. 2, sample RR-based affinity matrices are depicted. Two sets of experiments were conducted on the Beatles dataset. First, following the experimental setup employed in [1,3,6,7,12,14,17], the number of clusters (i.e., segments) K was kept constant and equal to 4. Second, for each music recording, the number of segments was estimated by (6). The optimal values of the various parameters (i.e., λ, τ) were determined by a grid search over 10 randomly selected music recordings of the dataset.

In order to compare fairly the RRMSA with the state-of-the-art music structure analysis methods, the segment labels are evaluated by employing the *pairwise F-measure* as in [3, 6, 7, 12, 14, 17]. The pairwise F-measure is a standard evaluation metric for clustering algorithms. It is defined as the harmonic mean of the pairwise precision and recall. The segmentation results and the reference segmentation (i.e., the ground truth) are handled at the time scale of beats. Let $\mathbb{F}_A$ be the set of identically labeled pairs of beats in a recording according to the music structure analysis algorithm and $\mathbb{F}_H$ be the set of identically labeled pairs in the human reference segmentation. The pairwise precision, PP, the pairwise recall, PR, and the pairwise F-measure, PF, are defined as:

$$PP = \frac{|\mathbb{F}_A \cap \mathbb{F}_H|}{|\mathbb{F}_A|}, \quad (7)$$

$$PR = \frac{|\mathbb{F}_A \cap \mathbb{F}_H|}{|\mathbb{F}_H|}, \quad (8)$$

$$PF = 2 \cdot \frac{PP \cdot PR}{PP + PR}, \quad (9)$$

where $|.|$ denotes the set cardinality.

4.2 Experimental Results

The segment-type labeling performance of the RRMSA on the Beatles dataset is summarized in Table 1 for a fixed number of segments (i.e., $K = 4$) as in [3, 6, 7, 12, 14, 17]. By inspecting Table 1, one can see that the ATMs are more suitable for music structure analysis than the MFCCs and the Chroma features. Furthermore, the latter two features lead to an undesirable over-segmentation of the music recordings. Similar findings were reported in [12]. The best result reported for segment-type labeling on the Beatles dataset is obtained here, when the RR-based affinity matrices of the MFCCs and the ATMs are combined. Interestingly to note that by employing cross-feature affinity matrices the average number of segments approaches 10 (i.e., the actual average number of segments according to the ground-truth), although no constraints have been enforced during clustering. In addition to the very promising performance of the RRMSA with respect to PF, it is worth mentioning that the RRMSA is fast. The average CPU time for the calculation of the RR-based affinity matrix is 0.858 CPU seconds.

Features	**Parameters**	PF	**Segments**
MFCCs	($\lambda = 0.3$)	0.54	37.1
Chroma	($\lambda = 0.1$)	0.57	36.7
ATMs	($\lambda = 0.1$)	**0.61**	6.1
MFCCs & Chroma	($\lambda = 0.3, 0.1$)	0.55	20.6
MFCCs & ATMs	($\lambda = 0.3, 0.1$)	**0.63**	7.1
Chroma & ATMs	($\lambda = 0.1, 0.1$)	0.60	8.1
MFCCs & Chroma & ATMs	($\lambda = 0.3, 0.1, 0.1$)	0.61	8.8

Table 1. Segment-type labeling performance of the RRMSA on the Beatles dataset with fixed $K = 4$.

The best result obtained by the RRMSA on the Beatles dataset with respect to PF (i.e., 0.63) outperforms the results obtained by the majority of the state-of-the-art music segmentation methods listed in Table 2 on the same dataset previously. The results were rounded down to the nearest second decimal digit. It is seen that the RRMSA admits the highest PF when the MFCCs and the ATMs are combined. Similarly, MFCCs combined with Chroma yielded the top PF in [3] and [6]. Similar conclusions were drawn in [13], when multiple audio features were combined. It is worth mentioning that the RRMSA does not involve any postprocessing based on music knowledge, such as eliminating too short segments or restricting the segment length to improve the accuracy of music segmentation. This is not the case for the methods in [7] and [14]. Furthermore, the RRMSA involves only one parameter in contrast to methods [3, 7, 14, 17], where the tuning of multiple parameters is needed.

The segment-type labeling performance of the RRMSA on the Beatles dataset, when K is estimated by (6), is reported in Table 3. Again, the use of the ATMs for music representation makes the RRMSA to achieve better performance than that when either the MFCCs or the Chroma features are used. By combining the ATMs and the MFCCs, the PF for the RRMSA reaches 0.60. In this case, the estimated average number of segments equals the actual av-

Reference	**Features**	PF
[3]	MFCCs & Chroma	**0.63**
[6]	MFCCs & Chroma	0.62
[17]	Chroma	0.60
[14]	MFCCs	0.60
[12]	ATMs	0.59
Method in [7] as evaluated in [14]	MPEG-7	0.58

Table 2. Segment-type labeling performance on the Beatles dataset obtained by state-of-the-art methods with fixed $K = 4$.

[4] http://www.dtic.upf.edu/ perfe/annotations/sections/license.html

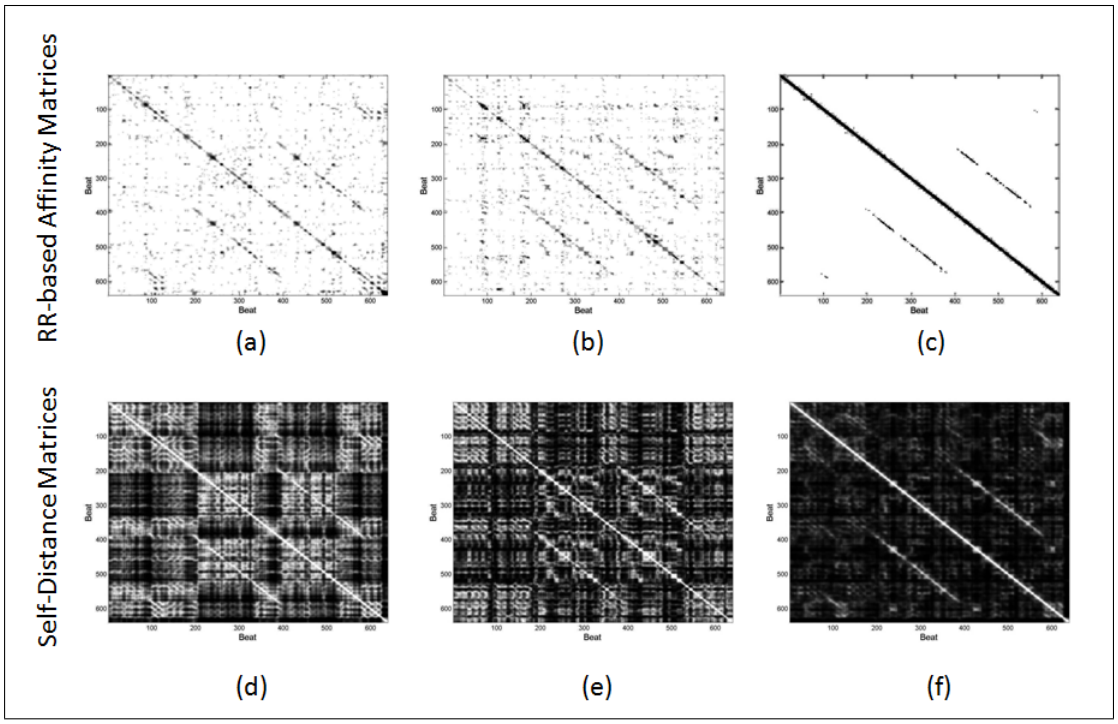

Figure 2. RR-based affinity and self-distance matrices of beat-synchronous feature vectors extracted from the Anna (Go to Him) by The Beatles when employing the MFCCs (a) and (d), the Chroma features (b) and (e), or the ATMs (c) and (f). The negative image of the affinity matrices is depicted. It is obvious that RR-based affinity matrices provide more clear and noise-free structural information than the self-distance matrices for all features.

erage number of segments according to the ground-truth (i.e., 10). This result indicates that it is possible to perform a robust unsupervised music structure analysis in a fully automatic setting.

Further details related to the estimation of K by employing various audio features and their combinations are shown in Table 4. The absolute error is defined as $|\bar{K} - K_g|$, where K_g is the actual number of segments based on the ground-truth. The prediction rate refers to the ratio of the number of music recordings where the number of segments was predicted correctly over the total number of music recordings in the dataset. If we consider the value $\bar{K} = K_g \pm 1$ as the correct number of predicted segments, then we obtain the Proximal Prediction Rate (PPR) (i.e., the last column in Table 4). The results presented in Table 4 indicate that the combination of the MFCCs and the ATMs yields the lowest absolute error, resulting to the highest prediction rate and thus the highest segmentation accuracy.

Features	**Parameters**	PF	**Segments**
MFCCs	$(\lambda = 0.3, \tau = 0.7)$	0.54	24.9
Chroma	$(\lambda = 0.1, \tau = 0.64)$	0.48	26.6
ATMs	$(\lambda = 0.1, \tau = 0.64)$	**0.59**	6.4
MFCCs + Chroma	$(\lambda = 0.3, 0.1, \tau = 0.7)$	0.59	12.2
MFCCs & ATMs	$(\lambda = 0.3, 0.1, \tau = 0.23)$	**0.60**	10.0
Chroma & ATMs	$(\lambda = 0.3, 0.1, \tau = 0.27)$	0.56	12.8
MFCCs & Chroma & ATMs	$(\lambda = 0.3, 0.1, \tau = 0.33)$	0.53	20.0

Table 3. Segment-type labeling performance of the RRMSA on the Beatles dataset for automatically estimated K.

Features	**Absolute Error**	**Prediction Rate (%)**	**PPR (%)**
MFCCs	**1.24**	**25.26**	**65.59**
Chroma	1.88	15.59	42.47
ATMs	1.72	18.81	52.68
MFCCs & Chroma	1.88	15.60	42.47
MFCCs & ATMs	**1.15**	**30.10**	**73.11**
Chroma& ATMs	1.43	22.58	61.82
MFCCs & Chroma & ATMs	1.22	26.34	67.20

Table 4. Accuracy of the estimation of the number of segments, K, on the Beatles dataset.

5. CONCLUSIONS

In this paper, a robust and fast method for music structure analysis (i.e., the RRMSA) has been proposed. In particular, the ridge regression representation of the MFCCs, the Chroma, and the ATMs have been used to derive affinity matrices, where the normalized cuts algorithm has been applied to obtain the music structure. Among the three features, the ATMs and the MFCCs have been proved the most powerful. By linearly combining the RR-based affinity matrices of the MFCCs and the ATMs and applying next the normalized cuts, state-of-the-art performance on the Beatles dataset has been reported for a fixed number of segments. Furthermore, an accurate method to estimate the number of segments in each music recording has been developed, enabling a fully automatic unsupervised music structure analysis.

APPENDIX: PROOF OF THEOREM 1

Let us denote by $\{\mathfrak{S}_1, \mathfrak{S}_2, \ldots, \mathfrak{S}_K\}$, a collection of K independent subspaces. The direct sum of a collection of K subspaces is denoted by $\oplus_{k=1}^{K}\mathfrak{S}_k$. Let $\mathbf{Z}$ be the unique minimizer of (1) and $\mathbf{D}$ be a block-diagonal matrix with elements $d_{ij} = z_{ij}$, if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same subspace (i.e., music segment here), and $d_{ij} = 0$ otherwise. We can define $\mathbf{Q} = \mathbf{Z} - \mathbf{D}$. Without loss of generality let us suppose that $\mathbf{x}_j$ belongs to the ith subspace, i.e., $\mathbf{x}_j = [\mathbf{XZ}]_j \in \mathfrak{S}_i$. We can write $\mathbf{Q}$ as the sum of two matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ whose supports are on disjoint subsets of indices, such that $[\mathbf{XQ}_1]_j \in \mathfrak{S}_i$ and $[\mathbf{XQ}_2]_j \in \oplus_{k\neq i}^{K}\mathfrak{S}_k$. We show that $\mathbf{Q}_2 = \mathbf{0}$. For the sake of contradiction, we assume that $\mathbf{Q}_2 \neq \mathbf{0}$. Since $\mathbf{Z} = \mathbf{D} + \mathbf{Q}_1 + \mathbf{Q}_2$, we have $\mathbf{x}_j = [\mathbf{XZ}]_j = [\mathbf{X}(\mathbf{D} + \mathbf{Q}_1)]_j + [\mathbf{XQ}_2]_j$. Since $\mathbf{x}_j \in \mathfrak{S}_i$ and $[\mathbf{X}(\mathbf{D} + \mathbf{Q}_1)]_j \in \mathfrak{S}_i$, by the independence of subspaces, $\mathfrak{S}_i \cap \oplus_{k\neq i}^{K}\mathfrak{S}_k = \{0\}$, we should have $[\mathbf{XQ}_2]_j = \mathbf{0}$.

But $[\mathbf{XQ}_2]_j = \mathbf{0}$ implies, $\mathbf{x}_j = [\mathbf{XZ}]_j = [\mathbf{X}(\mathbf{D}+\mathbf{Q}_1)]_j$ and hence $\mathbf{D} + \mathbf{Q}_1$ is feasible solution of (1). By the fact that the supports of Q_1 and Q_2 are disjoint subsets of indices and Lemma 1, $\|\mathbf{D} + \mathbf{Q}_1\|_F^2 \leq \|\mathbf{D} + \mathbf{Q}_1 + \mathbf{Q}_2\|_F^2 = \|\mathbf{Z}\|_F^2$. That is $\mathbf{D} + \mathbf{Q}_1$, is a feasible solution of (1) attaining a smaller Frobenius norm than $\|\mathbf{Z}\|_F^2$, which contradicts the optimality of $\mathbf{Z}$. Thus, $\mathbf{Q}_2 = \mathbf{0}$, meaning that only the blocks that correspond to vectors in the true subspaces are nonzero.

Acknowledgements

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heraclitus II. Investing in Knowledge Society through the European Social Fund.

6. REFERENCES

[1] L. Barrington, A. Chan, and G. Lanckriet. Modeling music as a dynamic texture. *IEEE Trans. Audio, Speech, and Language Processing*, 18(3):602–612, 2010.

[2] R. Bhatia and F. Kittaneh. Norm inequalities for partitioned operators and an application. *Math. Ann.*, 287(1):719–726, 1990.

[3] R. Chen and L. Ming. Music structural segmentation by combining harmonic and timbral information. In *Proc. 12th Int. Conf. Music Information Retrieval*, pages 477–482, Miami, USA, 2011.

[4] R. B. Dannenberg and M. Goto. Music structure analysis from acoustic signals. In D. Havelock, S. Kuwano, and M. Vorländer, editors, *Handbook of Signal Processing in Acoustics*, pages 305–331. Springer, New York, N.Y., USA, 2008.

[5] D. Ellis. Beat tracking by dynamic programming. *J. New Music Research*, 36(1):51–60, 2007.

[6] F. Kaiser and T. Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *Proc. 11th Int. Conf. Music Information Retrieval*, pages 429–434, Utrecht, The Netherlands, 2010.

[7] M. Levy and M. Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Trans. Audio, Speech, and Language Processing*, 16(2):318–326, 2008.

[8] G. Liu, Z. Lin, S. Yan, J. Sun, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011. arXiv:1010.2955v4 (preprint).

[9] R. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 1282–1285, Paris, France, 1982.

[10] M. Mauch, K. Noland, and S. Dixon. Using musical structure to enhance automatic chord transcription. In *Proc. 10th Int. Conf. Music Information Retrieval*, pages 231–236, Kobe, Japan, 2009.

[11] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. *IEEE Trans. Audio, Speech, and Language Technology*, 18(3):576–588, 2010.

[12] Y. Panagakis, C. Kotropoulos, and G. R. Arce. l1-graph based music structure analysis. In *Proc. 12th Int. Conf. Music Information Retrieval*, pages 495–500, Miami, USA, 2011.

[13] J. Paulus and A. Klapuri. Acoustic features for music piece structure analysis. In *Proc. 11th Int. Conf. Digital Audio Effects*, pages 309–312, Espoo, Finland, 2008.

[14] J. Paulus and A. Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Trans. Audio, Speech, and Language Processing*, 17(6):1159–1170, 2009.

[15] J. Paulus, M. Müller, and A. Klapuri. Audio-based music structure analysis. In *Proc. 11th Int. Conf. Music Information Retrieval*, pages 625–636, Utrecht, The Netherlands, 2010.

[16] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[17] R. Weiss and J. Bello. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization. In *Proc. 11th Int. Conf. Music Information Retrieval*, pages 123–128, Utrecht, The Netherlands, 2010.