A METHOD FOR IDENTIFYING REPETITION STRUCTURE IN MUSICAL AUDIO BASED ON TIME SERIES PREDICTION

Peter Foster, Anssi Klapuri and Simon Dixon

Centre for Digital Music
Queen Mary University of London
Mile End Road, London E1 4NS, UK
{peter.foster,anssi.klapuri,simon.dixon}@eecs.qmul.ac.uk

ABSTRACT

This paper investigates techniques for determining the repetition structure of musical audio. In particular, we consider the problem of determining segment similarity from the perspective of time series prediction, where we seek to quantify similarity in terms of pairwise predictability between segments. To this end, we propose a novel approach based on multivariate time series modelling of audio features.

Using chroma and MFCC features and based on the assumption that correct segment boundaries have been previously obtained, we evaluate the proposed approach against the Beatles dataset. We consider both Queen Mary and Tampere University versions of dataset annotations. We obtain a maximum pairwise F-score of 84%. Compared to a randomised baseline approach, this result corresponds to a performance improvement of 26 percentage points.

1. INTRODUCTION

In the field of music information retrieval, continued proliferation of music stored in digital audio format has been frequently cited as one impetus behind developing techniques for music signal analysis [1]. To the observer, the potential of developing techniques for automatically identifying musical attributes related to timbral, melodic, harmonic or rhythmic facets, is particularly apparent in the creation of large-scale music databases: Here, music signal analysis allows for novel applications in automated music search and recommendation systems which do not rely on manual annotation as a source of information.

In particular, the task of music structure analysis has recently received an increasing amount of attention in the literature [2]. In the context of music signal analysis, music structure analysis refers to identifying temporal structure at a time scale corresponding to the sectional form of a piece of music, where the notion of sectional form is applicable to a number of musics, including Western popular music. Music structure analysis lends itself to potential applications including music content navigation, fingerprinting and clustering [3]. Particularly relevant to music summarisation are applications such as chorus detection and thumbnailing, which may be facilitated by selecting a frequently occurring audio segment [4].

In this work, we propose a novel approach for detecting repetition in a piece of music, an important characteristic of temporal structure and a prerequisite for identifying sectional form. We motivate our approach considering the role of expectation in music listening [5]. In this view, we examine the problem of determining musical similarity in terms of prediction: We use a model of the statistics in musical time series, where the pairwise predictability of one time series relative to another allows us to quantify the probability of sequence repetition.

In music structure analysis, existing techniques for detecting repetition have widely featured self-similarity matrices [6], a form of recurrence plot obtained from a time-evolving sequence of feature vectors extracted from musical audio. In this approach, repeated musical sequences are manifest as diagonal lines in the self-similarity matrix, which may be identified using image processing techniques. An alternative class of approaches utilises clustering techniques to determine regions of homogeneity, where repeated sequences are identified by considering the mapping of sequences onto clusters [7]. In this formulation, techniques have been proposed based on hidden Markov models and histogram clustering [8].

In this work, the use of piecewise linear predictive models to aggregate statistics in sequences of feature vectors bears resemblance to the approach to music structure analysis proposed in [9], where sequences of feature vectors are parametrised using linear dynamical systems. With regard to estimating structural similarity using measures of pairwise predictability, [10] describes an evaluation of both linear and non-linear techniques for the problem of identifying 'cover songs', which may be defined as renditions of previously recorded music. Finally, note that multivariate time series modelling has been previously applied in diverse settings, including object boundary detection [11].

2. PROBLEM DEFINITION

Assume that we have a time series $V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N)$, corresponding to acoustic features observed at times $1, \ldots, N$. Each vector is a member of ℓ-dimensional feature space, with $\mathbf{v} \in \mathbb{R}^\ell$. The problem of inferring musical structure entails determining a set $S = \{s_1, \ldots, s_K\}$ of K segments. Each segment consists of a subset of observation times, so that for the kth segment we have $s_k = \{c_k + 1, \ldots, c_k + N_k\}$, with $s_k \in S$, $0 \leq c_k < N$ and $1 \leq c_k + N_k \leq N$. The set S is required to be a partition of the integers $\{1, \ldots, N\}$, so that segments are non-overlapping and the segmentation covers the entire piece. Segment boundaries are a description of a piece's sectional form and since intervals between segment boundaries are typically large, it follows that $|S| \ll N$. The process referred to as music structure analysis may be firstly viewed as the process of determining segment boundaries s_k. In a second process, segments are grouped according to

musical function: Assuming that a piece of music may be described in terms of a verse-chorus structure, a valid grouping might assign all segments corresponding to instances of the chorus to a single group. Formally, we can define a surjection $l : S \rightarrow \{1, \ldots, L\}$, with $L \leq K$. Then, the image of l yields information on which segments share identical function and which segments are distinct, in terms of sectional form. The set $G_n = \{s \in S : l(s) = n\}$ denotes all segments belonging to the nth group.

3. PROPOSED APPROACH

In this work, we assume that segment boundaries are given and that the task is to group identical segments. Our approach consists of three steps:

- Segment-wise predictive modelling of feature vector time series
- Accumulation of pairwise prediction accuracy between segments
- Segment clustering based on prediction accuracy

Let us consider with $\sigma : S \times S \rightarrow \mathbb{R}$ a distance function on pairs of segments. Intuitively, σ should express the amount of temporal structure shared between segments; it should discriminate strongly between pairs of segments obtained from the same group, versus segments obtained from different groups.

3.1. Predictive models

Our approach is based on predicting the sequence of feature vectors contained in segment s_r, with respect to a model estimated on segment s_q. We hereafter refer to this process as cross-prediction. In this view, cross-prediction accuracy quantifies pairwise similarity between segments. One means of predicting the sequence of feature vectors is to assume a vector autoregressive (VAR) model [12], where any given observation $\mathbf{v}_t$ is expressed as a linear combination of immediately preceding observations $\mathbf{v}_{t-M}, \mathbf{v}_{t-M+1}, \ldots, \mathbf{v}_{t-1}$, so that

$$\mathbf{v}_t = \sum_{m=1}^{M} \mathbf{A}_m \mathbf{v}_{t-m} + \mathbf{r}_t + \mathbf{c}. \tag{1}$$

In Equation 1, matrices $\mathbf{A}_m$ parametrise the dependence between $\mathbf{v}_t$ and its predecessor at lag m. The model order is specified by M and residuals $\mathbf{r}_t$ are assumed to have been generated by a multivariate Gaussian noise process with zero mean. The intercept vector $\mathbf{c}$ denotes the multivariate process mean.

We denote with $v_{t,n}$ the nth component of $\mathbf{v}_t$. Assuming independence between components, matrices $\mathbf{A}_m$ are diagonal. The VAR model is then equivalent to a component-wise autoregressive (AR) model,

$$v_{t,n} = \sum_{m=1}^{M} a_{m,n} v_{t-m,n} + r_{t,n} + c. \tag{2}$$

with $1 \leq n \leq \ell$, and where residuals r_{t_n} are assumed to have been generated by a univariate Gaussian noise process. The intercept c denotes the univariate process mean. We evaluate both VAR and component-wise AR models in our approach. In addition, we evaluate a separate model whose predictions are given by the multivariate process mean $\mathbf{c}$.

3.2. Modelling nonstationary time series

To account for nonstationarity in the time series, we window the time series and estimate an autoregressive model at each of the obtained window positions, as described in the following. Let us use $U_{i,q}$ to refer to the parameters of an VAR or component-wise AR model, disregarding the noise component, estimated on the ith subsequence formed when windowing observations within segment s_q according to some specified window length. We denote with n_q the number of window positions. The ith subsequence is demarcated by starting and ending positions $g_{i,q}$, $h_{i,q}$, where $1 \leq g_{i,q} < h_{i,q} \leq N$. We denote with $\mathcal{U}_q$ the sequence of models formed when considering all of n_q window positions in segment s_q,

$$\mathcal{U}_q = \left(U_{1,q}, U_{2,q}, \ldots, U_{n_q,q}\right). \tag{3}$$

Next, let us denote with $(\tilde{\mathbf{v}}_1^{i,q}, \tilde{\mathbf{v}}_2^{i,q}, \ldots, \tilde{\mathbf{v}}_N^{i,q})$ the sequence of features predicted by model $U_{i,q}$, as observations unfold in time. That is, $\tilde{\mathbf{v}}_t^{i,q}$ denotes the prediction given observations $\mathbf{v}_{t-M}, \mathbf{v}_{t-M+1}, \ldots, \mathbf{v}_{t-1}$ and given model $U_{i,q}$. To determine the cross-prediction accuracy with respect to model $U_{i,q}$, one possible approach might involve computing the sequence of squared errors

$$\epsilon_t^{i,q} = (\tilde{\mathbf{v}}_t^{i,q} - \mathbf{v}_t)^2. \tag{4}$$

However, if we consider the hypothetical case of a segment which exhibits high prediction error with respect to itself, it follows that its cross-prediction accuracy with respect to a segment consisting of identical features will be equivalent. Therefore, we might conclude that is necessary to place additional requirements on the utilised predictive model, or to further normalise prediction errors.

3.3. Autoregressive distance

Instead, we utilise the approach described in [13], which quantifies the structural similarity between two time series, using their estimated AR representations. The AR metric is based on the premise that for two AR processes defined by coefficient vectors $\mathbf{e}$, $\mathbf{f}$, identical starting conditions result in identical predictions being formed by both processes respectively, if $\mathbf{e} = \mathbf{f}$. We determine the dissimilarity between said AR processes by computing the squared AR distance $d^2(\mathbf{e}, \mathbf{f})$, defined as

$$d^2(\mathbf{e}, \mathbf{f}) = \|\mathbf{e} - \mathbf{f}\|^2 \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm. Note that in the formulation, we assume zero mean processes and disregard the variance of the error term. Applied to Equation 1, to determine the dissimilarity between two VAR models defined respectively by coefficients

$$\mathbf{E} = [\mathbf{E}_1, \ldots, \mathbf{E}_L] \; \mathbf{F} = [\mathbf{F}_1, \ldots, \mathbf{F}_M] \tag{6}$$

we compute $d^2_{\mathrm{VAR}}(\mathbf{E}, \mathbf{F}) = d^2(\mathrm{vec}(\mathbf{E}), \mathrm{vec}(\mathbf{F}))$, where $\mathrm{vec}(\cdot)$ denotes matrix vectorisation. We assume that model selection has been applied and that orders L, M therefore may differ. To accommodate the case where model orders differ, we perform zero padding prior to vectorisation. For the AR model of multivariate prediction, the distance vector $\mathbf{d}^2_{AR}(\mathbf{E}, \mathbf{F})$ is equivalent to applying Equation 5 component-

wise to AR models, as defined in Equation 2, and summing across components.

We compute the pairwise dissimilarity between segments s_q, s_r by interpolating between sequences of predictors. Denoting the number of multivariate predictors in each segment with n_q, n_r respectively, we average over VAR distances, so that the pairwise dissimilarity $\sigma_{\mathrm{VAR}}(s_q, s_r)$ is defined as

$$\sigma(s_q, s_r) = \frac{1}{K} \sum_{k=1}^{K} d^2_{\mathrm{VAR}} \left(U_{f(k,n_q),q}, U_{f(k,n_r),r} \right) \quad (7)$$

with $K = \max\{n_q, n_r\}$ and with $f(\cdot, \cdot)$ defined as

$$f(k, n) = \lceil k/K\, n \rceil. \quad (8)$$

Analogously, for the AR model of multivariate prediction, component-wise application of the AR distance results in a distance vector $\sigma_{\mathrm{AR}}(s_q, s_r)$. Finally, we separately compute and evaluate the average squared error between intercept terms $\sigma_\mu(s_q, s_r)$, which is equivalent to the average squared cross-prediction error when using the process mean to form predictions.

3.4. Segment clustering

Having obtained pairwise dissimilarity between segments, we apply the logistic function to the quantity $\sigma(s_q, s_r)$,

$$P(\sigma(s_q, s_r)) = \frac{1}{1 + \exp(-(\sigma(s_q, s_r)\alpha + \beta))} \quad (9)$$

where parameters α, β are obtained from a logistic regression. In the case of σ_{AR} and σ_μ, α is a row vector.

To determine segment identities, as proposed in [14] we interpret $P(\sigma(s_q, s_r))$ as the probability that segments s_q, s_r belong to the same group. Hereafter, we abbreviate the notation so that $P(\sigma_{i,j}) = P(\sigma(s_i, s_j))$. For each group we define a cluster C_m, such that $\bigcup_m C_m = \{1, \ldots, |S|\}$, $\bigcap_m C_m = \emptyset$. The set $T = \{C_1, \ldots, C_M\}$ defining a clustering is a partition whose log likelihood function is given by

$$P(T) = \frac{1}{2} \sum_{m=1}^{M} \sum_{i,j \in C_m, i \neq j} \log P(\sigma_{i,j}) + \sum_{j \notin C_m} \log(1 - P(\sigma_{i,j})) \quad (10)$$

whose value we seek to maximise. As the number of clusters M is not known a priori, we select the clustering which minimises Akaike's information criterion (AIC), defined as

$$\mathrm{AIC} = 2M - 2 \ln L_M \quad (11)$$

where L_M denotes the maximised likelihood of the clustering when assuming M clusters. Additionally, we consider Schwarz's Bayesian information criterion (BIC), defined as

$$\mathrm{BIC} = M \ln |S|^2 - 2 \ln L_M. \quad (12)$$

4. RESULTS

We evaluate the proposed approach based on a dataset consisting of 180 mono audio tracks of all studio albums by the Beatles. Each track is sampled at 44.1kHz.

4.1. Feature extraction

To incorporate information of harmonic musical content, we extract chroma features using the method and implementation described in [15]. The chroma features are obtained in two steps. In the first step, a log-spaced multirate filter bank is applied whose band centre frequencies are based on pitches in the chromatic scale between A0 and C8, assuming equal-tempered tuning and shifting centre frequencies according to estimated tuning deviation. A spectro-temporal representation of signal power is then formed by applying a sliding window of 200ms length and 50% overlap, resulting in a frame rate of 10Hz. In the second step, frame components corresponding to the same pitch class are summed, where logarithmic compression is applied to the range of pitches, to account for approximate human perception of sound intensity. The resulting 12-component feature vectors are then normalised, where components of low energy frames are beforehand replaced with uniform values to avoid noise-like behaviour in non-melodic portions of the music signal.

As a descriptor of musical timbre, we extract MFCCs using the method and implementation described in [16]. The MFCC features are extracted using an FFT window size of 4410 samples and 50% overlap. The applied Mel scale is approximated using window centre frequencies spaced $66\frac{2}{3}$Hz apart, followed by 27 logarithmically spaced centre frequencies, where in the latter case the factor relating consecutive centre frequencies is approximately 1.07. The lowest frequency window is centred at $133\frac{1}{3}$Hz. We disregard the zeroth coefficient and retain the first 12 remaining coefficients.

4.2. Prediction

The estimation of predictor sequences is based on a window size in the range of 25 to 150 frames and a 25 frame hop size. Prior to AR estimation, we apply principal components analysis to the entire time series as a means of decorrelating feature vectors, yielding a further 12-component time series. Prior to VAR and AR predictor estimation, time series contained at each window position are centred to have zero mean. Model orders are determined using the BIC. Model parameters for AR models are estimated using the Levinson-Durbin method, whereas VAR models are estimated according to the method and implementation described in [17].

4.3. Clustering

We acquire pairwise segment similarities using the approach described in Section 3.3. To determine segment groups, the clustering problem is expressed as its graph theoretic dual. We estimate the maximally likely clustering using the normalised cut approach described in [18]. Logistic regression is applied to segment similarity data using a 10-fold cross-validation approach.

4.4. Performance statistics

To evaluate the accuracy of the proposed approach, we utilise the pairwise precision, recall and F-score procedure as proposed in [8]. Given a song, we denote with R_a the set of identically labelled frame pairs in the annotation, correspondingly we denote with R_b the conjectured set of identically labelled frame pairs. The pairwise precision PP and pairwise

recall PR are defined as

$$PP = \frac{|R_a \cap R_b|}{|R_b|} \quad (13)$$

$$PR = \frac{|R_a \cap R_b|}{|R_a|}. \quad (14)$$

The pairwise F-score is defined as the harmonic mean of PR and PF,

$$PF = 2\,\frac{PP\ \ PR}{PP + PR}. \quad (15)$$

We evaluate using two separate versions of dataset annotations, available from Tampere University [1] and Queen Mary University of London (QMUL) [2]. Fig. 1 displays a histogram of segment labels contained in the annotation datasets.

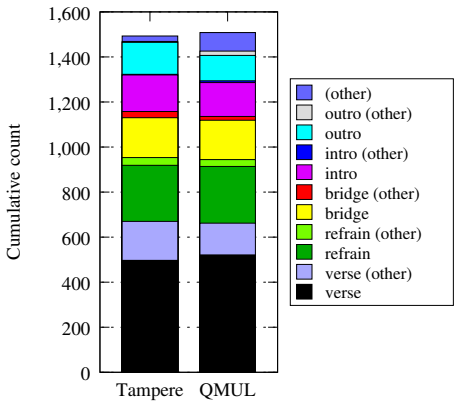


Fig. 1. Stacked histogram of segment labels defined in Tampere and QMUL datasets.

4.5. Performance

Figures 2, 3 display plots of pairwise F-scores, for combinations of data sets, audio features, model selection methods and for each of the prediction methods described in Section 3.1: Labels 'AR', 'VAR', 'Mean' correspond to AR, VAR and process mean prediction approaches, respectively. As a baseline, we evaluate an approach where clustering is applied to similarities sampled from a normal distribution, labelled 'Random' in the figures. The aforementioned approaches are combined with AIC and BIC model selection, as indicated by label suffixes.

We observe that for the QMUL annotations (Fig. 2), the AR and 'Mean' approaches consistently outperform the baseline approach. This result holds for all evaluated window sizes. For both the AR and VAR approaches, we observe that the BIC consistently outperforms the AIC as a means of model selection. In the case of chroma features and comparing the best-performing AR and VAR approaches, we observe no significant difference in performance (Figs. 2 a,c). However, in the case of MFCC features, we observe that the best-performing AR approach significantly outperforms the best-performing VAR approach by 6% and 7%, for AIC and BIC, respectively (Figs. 2 b,d). In overall comparison to the AR approach, the VAR approach appears more sensitive to the window size, exhibiting larger amounts of variance in performance. For the Tampere annotations, this behaviour is particularly apparent (Fig. 3 c).

Furthermore, we observe that the 'Mean' approach outperforms both AR and VAR approaches by an average of 11 and 19 percentage points respectively, averaged over considered window sizes and data sets. Considering that the autoregressive and mean-value approaches incorporate different statistics on time series, we conjecture that combining both aforementioned approaches might result in an improvement over the case of using the 'Mean' approach alone. In initial evaluations, a linear combination of segment similarities did not yield a performance gain beyond the obtained maximum 'Mean' performance. To this end, we are therefore currently evaluating the use of alternative ensemble techniques.

Table 1 displays a summary of performance for each of the approaches, maximised over window sizes.

Approach	Maximum pairwise F-Score (%)
Mean,Chroma	82.54
Mean,MFCC	83.78
AR,Chroma	70.63
AR,MFCC	74.11
VAR,Chroma	68.24
VAR,MFCC	68.28
Random	57.74

Table 1. Summary of results. See main text for a description of row labels.

5. CONCLUSIONS AND FURTHER WORK

In this work, we have considered the problem of determining the repetition structure of musical audio. We have considered how predictability might be used as a measure of similarity between sequences. We have proposed a novel method for detecting the repetition structure of musical audio, based on multivariate time series prediction. The proposed approach requires no prior knowledge of the number of segment groups, relying instead on model selection for clustering segments.

Evaluated against two versions of dataset annotations, using chroma and MFCC features, the obtained results suggest that the proposed method is a viable approach for the aforementioned problem: Although employing a restricted amount of domain knowledge, the proposed method improves significantly over the baseline.

Considering the obtained results, we plan to extend the present evaluation to cater for the case where segment boundaries are obtained heuristically. In addition, we plan to incorporate additional domain knowledge in the form of segment duration models. Furthermore, we aim to evaluate a broader range of predictors, including non-linear prediction techniques.

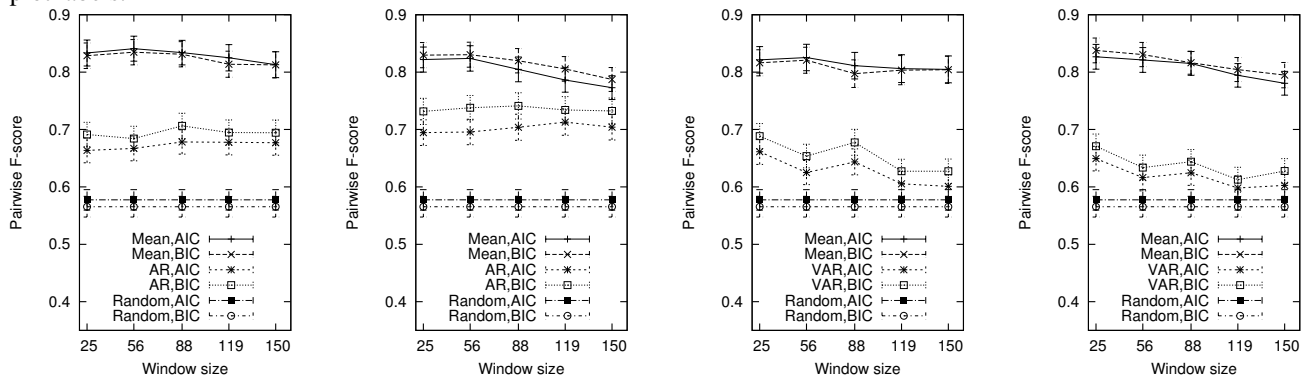
6. ACKNOWLEDGEMENTS

This work was supported by funding from the Engineering and Physical Sciences Research Council (United Kingdom). In addition, we thank the anonymous reviewers for their comments.

[1] http://www.cs.tut.fi/sgn/arg/paulus/beatles_sections_TUT.zip

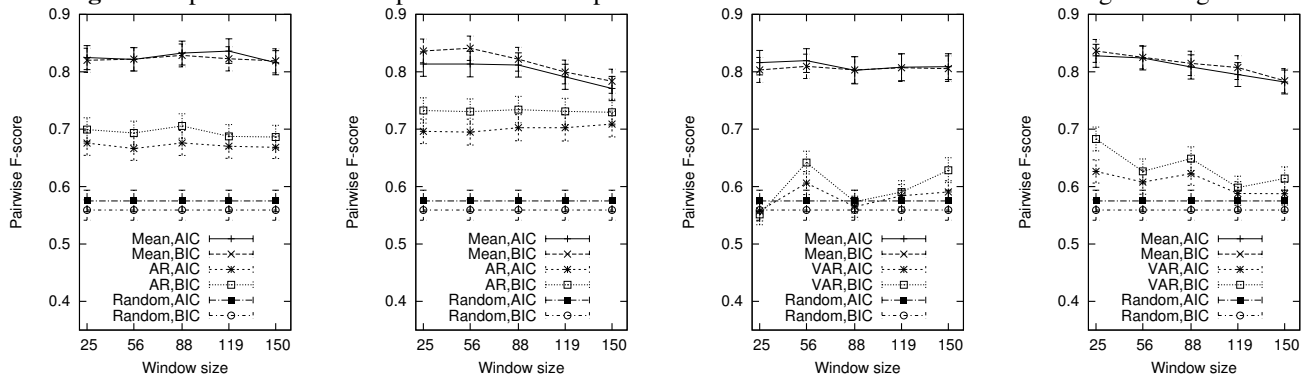
[2] http://isophonics.net/content/reference-annotations-beatles

Fig. 2. QMUL annotation set performance. Error bars correspond to 95% confidence intervals. See main text for a description of plot labels.



(a) Chroma, AR model (b) MFCC, AR model (c) Chroma, VAR model (d) MFCC, VAR model

Fig. 3. Tampere annotation set performance. The plot labels and error bars have the same meaning as in Fig. 2.



(a) Chroma, AR model (b) MFCC, AR model (c) Chroma, VAR model (d) MFCC, VAR model

7. REFERENCES

[1] J.J. Aucouturier, F. Pachet, and M. Sandler, "The way it sounds: Timbre models for analysis and retrieval of music signals," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1028–1035, 2005.

[2] J. Paulus, M. Müller, and A. Klapuri, "State of the art report: Audio-based music structure analysis," in *Proc. 11th Intern. Society for Music Information Retrieval Conf.*, 2010, pp. 625–36.

[3] B.S. Ong, *Structural Analysis and Segmentation of Music Signals*, Ph.D. thesis, University Pompeu Fabra, Barcelona, Spain, 2007.

[4] M.A. Bartsch and G.H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[5] D.B. Huron, *Sweet anticipation: Music and the psychology of expectation*, The MIT Press, 2006.

[6] J. Foote, "Visualizing music and audio using self-similarity," in *Proc. Seventh ACM Intern. Conf. on Multimedia (Part 1)*. ACM, 1999, pp. 77–80.

[7] G. Peeters, "Deriving musical structures from signal analysis for music audio summary generation:"sequence" and "state" approach," *Lecture Notes in Computer Science*, pp. 169–185, 2004.

[8] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[9] L. Barrington, A.B. Chan, and G. Lanckriet, "Modeling music as a dynamic texture," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 602–612, 2010.

[10] J. Serra, H. Kantz, X. Serra, and R.G. Andrzejak, "Predictability of music descriptor time series and its application to cover song detection," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 514–525, 2012.

[11] T. Quatieri, "Object detection by two-dimensional linear prediction," in *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 1983, vol. 8, pp. 108–111.

[12] H. Lütkepohl, *New introduction to multiple time series analysis*, Springer, 2005.

[13] D. Piccolo, "A distance measure for classifying ARIMA models," *Journal of Time Series Analysis*, vol. 11, no. 2, pp. 153–164, 1990.

[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*, vol. 17, no. 6, pp. 1159–1170, 2009.

[15] M. Müller and S. Ewert, "Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features," in *Proc. 12th Intern. Society for Music Information Retrieval Conf.*, 2011.

[16] M. Slaney, *Auditory toolbox version 2*, Interval Research Corporation, 1998.

[17] A. Neumaier and T. Schneider, "Estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM Trans. Math. Software (TOMS)*, vol. 27, no. 1, pp. 27–57, 2001.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.