STATISTICAL MUSIC STRUCTURE ANALYSIS BASED ON A HOMOGENEITY-, REPETITIVENESS-, AND REGULARITY-AWARE HIERARCHICAL HIDDEN SEMI-MARKOV MODEL

Go Shibata **Ryo Nishikimi** **Eita Nakamura** **Kazuyoshi Yoshii**
Graduate School of Informatics, Kyoto University, Japan
{gshibata, nishikimi, enakamura, yoshii}@sap.ist.i.kyoto-u.ac.jp

ABSTRACT

This paper describes a music structure analysis method that splits music audio signals into meaningful segments such as musical sections and clusters them. In this task, how to model the four fundamental aspects of musical sections, *i.e.*, homogeneity, repetitiveness, novelty, and regularity, in a unified way is still an open problem. Here we propose a solid statistical approach based on a homogeneity-, repetitiveness-, and regularity-aware hierarchical hidden semi-Markov model. The higher-level semi-Markov chain represents a sequence of sections that tend to have *regularly* spaced boundaries. The timbral features in each section are assumed to follow emission distributions that are *homogeneous* over time. The lower-level left-to-right Markov chain in each section represents a chord sequence whose sequential order is constrained to be a *repetition* of a chord sequence in another section of the same cluster. The whole model can be trained unsupervisedly based on Bayesian sparse learning where unnecessary sections automatically degenerate. The proposed method outperformed representative methods in segmentation and clustering accuracies with estimated sections having similar statistical properties as the ground truth data.

1. INTRODUCTION

Music structure analysis is a long-standing research topic [1] because detection of meaningful segments called musical sections (*e.g.*, intro, verse, bridge, and chorus sections in popular music) from music audio signals forms a basis of music information retrieval (MIR). In general, music structure analysis involves a *segmentation* step that splits music signals into sections [2–9], a *clustering* step that categorizes such sections into several classes [10–18], and a *labeling* step that gives each section a concrete label such as "verse A", "verse B", or "chorus" [19–21]. We here tackle the segmentation and clustering for popular music.

In previous studies, sections of popular music have been characterized in three aspects, *i.e.*, *homogeneity* referring to the intra-section characteristics and *repetitiveness* and *novelty* referring to the inter-section relationships [1]. More specifically, homogeneity means that musical characteristics (*e.g.*, timbral features such as mel-frequency cepstrum coefficients (MFCCs)) are consistent within a section. Repetitiveness means that the same sequence of some musical elements (*e.g.*, chroma features and chord progressions) of a section is repeated in sections of the same class. Novelty means that musical characteristics change abruptly at a boundary between sections. In addition, *regularity* of section durations has often been focused on [8–10] because there are typical lengths such as four or eight measures in popular music.

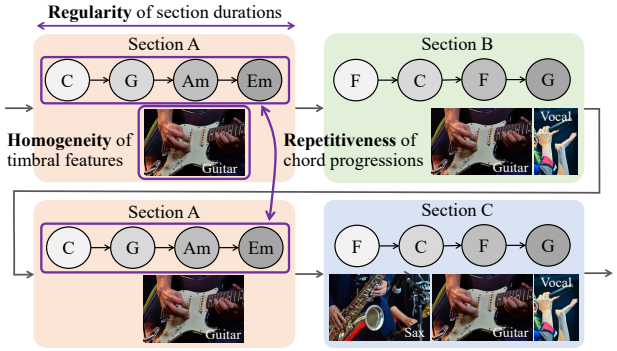

Figure 1. Music structure analysis based on homogeneity of timbral features, repetitiveness of chord progressions, and regularity of section durations.

Most studies on music structure analysis, however, have focused on only one of the above aspects or consider some of them in a separate and/or ad-hoc manner, as reviewed in Section 2. Joint computational modeling of these four aspects is thus the central issue in music structure analysis. Instead of manually designing segmentation and clustering criteria based on these aspects, we pursue a statistical approach to *data-driven* music structure analysis.

In this study, we propose a statistical music structure analysis method that simultaneously deals with the homogeneity, repetitiveness, and regularity of musical sections in a probabilistic framework (Fig. 1). We formulate a unified probabilistic model called a hierarchical hidden semi-Markov model (HHSMM) that represents the hierarchical generative process of musical sections, chord sequences, and music audio signals (timbral features and chroma features). This model has two sequences of latent states in a hierarchical manner. The upper-level sequence

represents a series of section classes following a semi-Markov model with explicit regularity (duration) modeling and the lower-level sequence represents chords following a section-conditioned left-to-right Markov model. To represent the intra-section homogeneity of timbral features, the MFCCs of a section are assumed to be generated from an upper-level state corresponding to the section. To represent the inter-section repetitiveness of chord progressions, the chroma features of a section are assumed to be sequentially generated from lower-level states. Given a music audio signal as observed data, the whole model can be trained unsupervisedly using Gibbs sampling and Viterbi training, where unnecessary sections are automatically degenerated during the Bayesian sparse learning.

The main contribution of this study is to propose a solid Bayesian approach to music signal analysis based on a fully *generative* model that can deal with the homogeneity, repetitiveness, and regularity of sections in a unified way. This enables *unsupervised* learning unlike another statistical approach based on deep *discriminative* models [6–8] that require section annotations for supervised training. Since these two approaches have a mutually complementary relationship, our results open up a door to deep Bayesian integration of discriminative and generative models in a variational autoencoding framework (audio signal $\rightarrow$ sections $\rightarrow$ audio signal) [22] for further improvement.

Another important contribution is to investigate the statistical characteristics of musical sections estimated by the proposed method in comparison with representative methods. Our method is shown to be able to yield distributions of section durations, of the numbers of section classes used for representing music audio signals, and of the metrical positions of section boundaries much more similar to those of the ground-truth data than the other methods.

2. RELATED WORK

Homogeneity, repetitiveness, novelty, and regularity have been considered to be important for music structure analysis. The most standard approach to music structure analysis is to use the self-similarity matrix (SSM) of acoustic features such as chroma features and MFCCs, whose element represents the acoustic similarity between two time frames (Fig. 2). In an SSM, the homogeneity, repetitiveness, novelty, and regularity are observed as block-diagonal structure, short stripes parallel to the diagonal line, grid patterns, and grid intervals, respectively. One or some of these four aspects have been used for segmentation and clustering tasks in music structure analysis.

2.1 Segmentation

Foote [2] proposed a novelty-based method that detects peaks from a time-varying novelty curve obtained by shifting and convoluting a checkerboard kernel along the diagonal elements of an SSM. Jensen [3] attempted to find section boundaries that minimize a homogeneity- and novelty-based cost function. While Goto [19] originally proposed a lag SSM in which repetitions appear not as stripes but as vertical lines and calculated a novelty curve over

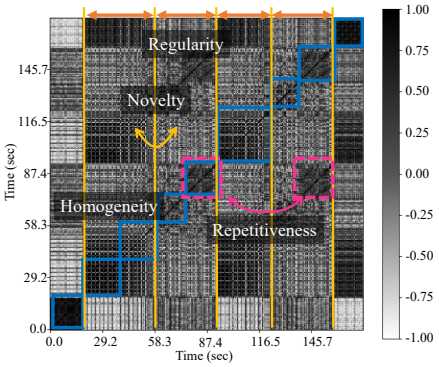

Figure 2. Self-similarity matrix (SSM) of MFCC features (part of RWC-MDB-P-2001 No. 25).

time lags, Serrà [4] proposed another novelty curve over time frames. Peeters and Bisot [5] successfully integrated these two methods [4, 19] for better segmentation. Ullrich *et al.* [6] pioneered a supervised approach based on a convolutional neural network (CNN), which was extended to deal with both coarse- and fine-level boundary annotations [7]. Sargent *et al.* [9] pointed out the effectiveness of focusing on the regularity to favor structural segments of comparable size. Maezawa [8] developed a long short-term memory (LSTM) network with homogeneity-, repetitiveness-, novelty-, and regularity-based cost functions. In this study we take an *unsupervised* approach based on a homogeneity-, repetitiveness-, and regularity-based *generative* model. To keep the model simple, incorporation the aspect of novelty is left for future work.

2.2 Clustering

Cooper *et al.* [12] sequentially performed music segmentation [2] and section clustering based on intra- and inter-section statistical characteristics. Goodwin *et al.* [13] attempted to efficiently detect off-diagonal stripes in an SSM as repetitions using dynamic programming. To deal with repetitiveness and homogeneity, Grohganz *et al.* [14] converted a repetitiveness-dominant SSM with off-diagonal stripes into a homogeneity-dominant SSM with block-diagonal structure. Nieto *et al.* [15] used a convex variant of non-negative matrix factorization for section segmentation and clustering. McFee *et al.* [16] proposed a method that encodes repetitive structures into a graph and performs spectral clustering for graph partitioning.

Several studies took a statistical approach based on generative models for joint segmentation and clustering. Aucouturier *et al.* [11] used a standard hidden Markov model (HMM). Ren *et al.* [17] proposed a nonparametric Bayesian extension of an HMM that can estimate an appropriate number of sections. Barrington *et al.* [18] proposed a nonparametric Bayesian extension of a switching linear dynamical system (LDS) that also has the ability of automatic model complexity control. While these methods mainly focused on the homogeneity, our method simultaneously considers the homogeneity, repetitiveness, and regularity and can incorporate prior knowledge about the statistical characteristics of sections (*e.g.*, durations and metrical positions) in a data-driven manner.

3. PROPOSED METHOD

This section describes the proposed statistical method for music structure analysis.

3.1 Problem Specification

Our problem is formulated as follows:

Input: A chroma feature sequence $\mathbf{X}^c = \mathbf{x}^c_{1:B} \in \mathbb{R}^{B\times 12}$ and an MFCC sequence $\mathbf{X}^m = \mathbf{x}^m_{1:B} \in \mathbb{R}^{B\times 12}$ obtained from a given music audio signal. They are computed in units of beats estimated by a beat tracking method [23].
Output: Boundaries and classes of sections.

Here, B is the number of beats (quarter notes) and a subscript $\bigcirc_{a:b}$ represents the sequence $(\bigcirc_a, \ldots, \bigcirc_b)$.

We use 12-dimensional chroma features obtained by aggregating all spectral information of each pitch class into a single element. We also use 12-dimensional MFCCs as timbral features.

3.2 Model Formulation

The proposed model consists of two-level hierarchical Markov chains and an acoustic model as shown in Fig. 3. The upper-level Markov chain describes the section-level structure (*i.e.*, section classes and durations) and the lower-level Markov chain describes the internal structure of each section. The states of these Markov chains are latent variables that represent abstract musical structure. The acoustic model connects the abstract structure with the observed musical features (chroma vectors and MFCCs).

3.2.1 Upper-Level Markov Chain

The upper-level Markov chain is an ergodic semi-Markov model. The sequence of section classes $\mathbf{Z} = z_{1:T}$ ($z_\tau \in \{1, \ldots, N_Z\}$) and their durations $\mathbf{D} = d_{1:T}$ ($d_\tau \in \{1, \ldots, N_D\}$) are generated by the model, where T is the number of sections, N_Z is the number of distinct section classes, and N_D is the maximum duration of a section. The generative process is described as follows:

$$p(z_1, d_1) = \rho_{z_1}\psi_{d_1}, \tag{1}$$

$$p(z_\tau, d_\tau | z_{\tau-1}, d_{\tau-1}) = \pi_{z_{\tau-1}z_\tau}\psi_{d_\tau}, \tag{2}$$

where ρ_z and $\pi_{zz'}$ are the initial and transition probabilities of section classes and ψ_d is the duration probability.

3.2.2 Lower-Level Markov Chain

The lower-level Markov chain is a left-to-right Markov model with N_K states. The state sequence of this model describes the internal structure of a section corresponding to the chord progression, where each state is expected to correspond to a chord. We consider such a Markov chain for each section class. The model continues state transitions for each beat from the start time of a section until its duration passes. The state sequence $\mathbf{K}_\tau = k_{\tau,1:d_\tau}$ ($k_{\tau,t} \in \{1, \ldots, N_K\}$) is generated as follows:

$$p(k_{\tau,t} | z_\tau, k_{\tau,t-1}) = \phi^{(z_\tau)}_{k_{\tau,t-1}k_{\tau,t}}, \tag{3}$$

where z_τ and d_τ are the corresponding section class and duration, and $\phi^{(z_\tau)}_{kk'}$ is the transition probability from state k to state k'.

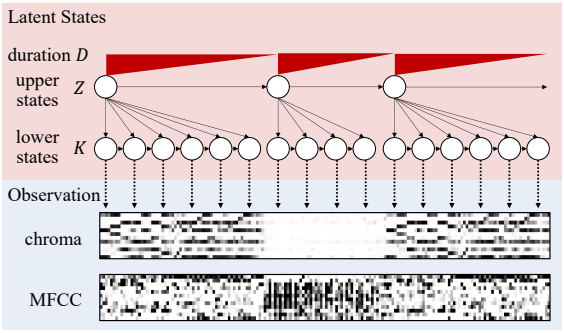

Figure 3. The proposed generative model.

The left-to-right Markov model meets a condition that the initial state has $k_{\tau,1} = 1$ and $k_{\tau,t_1} \le k_{\tau,t_2}$ for $t_1 < t_2$. We introduce a hyperparameter σ that describes the maximum number of states that may be skipped in a transition; a transition from state k to state $k+\sigma$ is possible but a larger skip is forbidden. In this way, the model can describe repetitions with some variations, which is another important aspect of musical structure. $\mathbf{K}$ denotes $\mathbf{K}_{1:T}$.

3.2.3 Acoustic Model

The acoustic model describes the generative processes of the chroma features $\mathbf{x}^c_b \in \mathbb{R}^{12}$ and MFCCs $\mathbf{x}^m_b \in \mathbb{R}^{12}$ by using output probabilities that are defined conditionally on the section classes $\mathbf{Z}$ and their internal states $\mathbf{K}$. The repetitiveness is represented by applying the same set of output probabilities to all sections of the same class. The output probabilities of chroma features $\chi^c_{z,k}$ depend on both $\mathbf{Z}$ and $\mathbf{K}$ to represent the sequential structure of chord progressions. To capture the homogeneity of timbre characteristics of each section, the output probabilities of MFCCs χ^m_z are assumed to depend only on $\mathbf{Z}$. Thus we have

$$p(\mathbf{x}^c_b, \mathbf{x}^m_b) = \chi^c_{z_b,k_b}(\mathbf{x}^c_b)\chi^m_{z_b}(\mathbf{x}^m_b), \tag{4}$$

where z_b and k_b are the section class and the internal state at beat b, respectively. The output probabilities are described as multivariate Gaussian distributions:

$$\chi^c_{z,k}(\mathbf{x}^c_b) = \mathcal{N}(\mathbf{x}^c_b | \boldsymbol{\mu}^c_{z,k}, (\boldsymbol{\Lambda}^c_{z,k})^{-1}), \tag{5}$$

$$\chi^m_z(\mathbf{x}^m_b) = \mathcal{N}(\mathbf{x}^m_b | \boldsymbol{\mu}^m_z, (\boldsymbol{\Lambda}^m_z)^{-1}), \tag{6}$$

where $\boldsymbol{\mu}^c_{z,k}$ and $\boldsymbol{\Lambda}^c_{z,k}$ are the mean and precision for the chroma features, and $\boldsymbol{\mu}^m_z$ and $\boldsymbol{\Lambda}^m_z$ are defined similarly.

3.2.4 Prior Distributions

To find an effective number of distinct section classes, we formulate a Bayesian HHSMM by putting conjugate prior distributions. We put Dirichlet prior distributions for the categorical distributions as follows:

$$\boldsymbol{\rho} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\rho}}), \tag{7}$$

$$\boldsymbol{\psi} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\psi}}), \tag{8}$$

$$\boldsymbol{\pi}_z \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\pi}}), \tag{9}$$

$$\boldsymbol{\phi}^{(z)}_k \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\phi}}), \tag{10}$$

where $\boldsymbol{\rho} = \rho_{1:N_Z}$, $\boldsymbol{\psi} = \psi_{1:N_D}$, $\boldsymbol{\pi}_z = \pi_{z(1:N_Z)}$, $\boldsymbol{\phi}^{(z)}_k = \phi^{(z)}_{k(1:N_K)}$, and $\mathbf{a}^{\boldsymbol{\rho}}$, $\mathbf{a}^{\boldsymbol{\psi}}$, $\mathbf{a}^{\boldsymbol{\pi}}$, and $\mathbf{a}^{\boldsymbol{\phi}}$ are Dirichlet parameters. When these parameters are small, the transition probabilities of section classes become sparse. The model can

thus capture repetitions regardless of small acoustic variations and remove unnecessary section classes.

Since section durations tend to be the integer multiples of four measures in popular music (see Fig. 4), such a statistical tendency can be incorporated in the prior distribution. Specifically, we use as $\mathbf{a}^{\boldsymbol{\psi}}$ the empirical distribution of section durations $\mathbf{a}^{\boldsymbol{\psi}}_{\mathrm{emp}}$ multiplied by a constant factor. Since the structure of section classes is quite different over individual musical pieces, we put uniform Dirichlet prior distributions for their transition probabilities.

Finally, we put Gaussian-Wishart prior distributions on multivariate normal distributions as follows:

$$\boldsymbol{\mu}^c_{z,k}, \boldsymbol{\Lambda}^c_{z,k} \sim \mathcal{N}(\boldsymbol{\mu}^c_{z,k}|\mathbf{m}^c_0, (\beta^c_0\boldsymbol{\Lambda}^c_{z,k})^{-1})\, \mathcal{W}(\boldsymbol{\Lambda}^c_{z,k}|\mathbf{W}^c_0, \nu^c_0),$$
$$\boldsymbol{\mu}^m_z, \boldsymbol{\Lambda}^m_z \sim \mathcal{N}(\boldsymbol{\mu}^m_z|\mathbf{m}^m_0, (\beta^m_0\boldsymbol{\Lambda}^m_z)^{-1})\, \mathcal{W}(\boldsymbol{\Lambda}^m_z|\mathbf{W}^m_0, \nu^m_0),$$

where $\mathbf{m}^c_0$, β^c_0, $\mathbf{W}^c_0$, ν^c_0, $\mathbf{m}^m_0$, β^m_0, $\mathbf{W}^m_0$, and ν^m_0 are hyperparameters.

3.3 Bayesian Learning

Letting $\boldsymbol{\Theta} = \{\boldsymbol{\rho}, \boldsymbol{\psi}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$, we aim to calculate the posterior distribution $p(\mathbf{Z}, \mathbf{D}, \mathbf{K}, \boldsymbol{\Theta}|\mathbf{X}^c, \mathbf{X}^m)$. Since this is analytically intractable, we use the Gibbs sampling method. We first sample the latent variables $\mathbf{Z}$, $\mathbf{D}$, and $\mathbf{K}$ from the distribution $p(\mathbf{Z}, \mathbf{D}, \mathbf{K}|\boldsymbol{\Theta}, \mathbf{X}^c, \mathbf{X}^m)$ and we then sample the model parameters $\boldsymbol{\Theta}$ from the distribution $p(\boldsymbol{\Theta}|\mathbf{Z}, \mathbf{D}, \mathbf{K}, \mathbf{X}^c, \mathbf{X}^m)$. Iterating this process, we obtain samples from the true posterior distribution.

3.3.1 Sampling Latent Variables

We use the forward filtering-backward sampling algorithm for sampling the upper- and lower-level latent variables $\mathbf{Z}$, $\mathbf{D}$, and $\mathbf{K}$. We introduce variables z_b and d_b that denote the class and duration of a section starting at beat $b-d_b+1$ and ending at beat b. We also define the marginalized output probabilities for this section $\omega_{z_b}(\mathbf{x}^c_{b-d_b+1:b}, \mathbf{x}^m_{b-d_b+1:b})$, which can be calculated by the forward algorithm for the lower-level Markov chain.

In the forward filtering step for the upper-level model, we initialize and update the forward variables $\alpha_b(z_b, d_b) = p(z_b, d_b, \mathbf{x}^c_{1:b}, \mathbf{x}^m_{1:b})$ as follows:

$$\alpha_b(z_b, d_b = b) = \rho_{z_b}\psi_{d_b}\omega_{z_b}(\mathbf{x}^c_{1:b}, \mathbf{x}^m_{1:b}), \tag{11}$$

$$\alpha_b(z_b, d_b) \tag{12}$$
$$= \sum_{z', d'} \alpha_{b-d_b}(z', d')\pi_{z'z_b}\psi_{d_b}\omega_{z_b}(\mathbf{x}^c_{b-d_b+1:b}, \mathbf{x}^m_{b-d_b+1:b}).$$

In the backward sampling step, the latent variables $\mathbf{Z}$ and $\mathbf{D}$ are sequentially sampled in the reverse order:

$$p(z_B, d_B|\mathbf{X}^c, \mathbf{X}^m) \propto \alpha_B(z_B, d_B). \tag{13}$$

When variables z_b and d_b are already sampled, the variables $z_{b'}$ and $d_{b'}$ at beat $b' = b - d_b$ are sampled according to the probability

$$p(z_{b'}, d_{b'}|z_{b:B}, d_{b:B}, \mathbf{X}^c, \mathbf{X}^m) \propto \alpha_{b'}(z_{b'}, d_{b'})\pi_{z_{b'}z_b}. \tag{14}$$

Next, the latent variables $\mathbf{K}$ are sampled using the sampled $\mathbf{Z}$ and $\mathbf{D}$. Each set of variables $\mathbf{K}_\tau$ is sampled by forward filtering-backward sampling for the lower-level model of section class z_τ. Here we use a beat index $t \in \{1, \ldots, d_\tau\}$ considered in relative to the section boundary. In the forward filtering step, we calculate the probabilities $\zeta_{\tau,k_{\tau,t}}$ recursively as follows:

$$\zeta_{\tau,k_{\tau,1}} = p(k_{\tau,1}, \mathbf{x}^c_1, \mathbf{x}^m_1|z_\tau, d_\tau)$$
$$= \delta_{k_{\tau,1}1}\chi^c_{z_\tau,1}(\mathbf{x}^c_1)\chi^m_{z_\tau}(\mathbf{x}^m_1), \tag{15}$$

$$\zeta_{\tau,k_{\tau,t}} = p(k_{\tau,t}, \mathbf{x}^c_{1:t}, \mathbf{x}^m_{1:t}|z_\tau, d_\tau) \tag{16}$$
$$= \left(\sum_{k_{\tau,t-1}} \zeta_{\tau,k_{\tau,t-1}}\phi^{(z_\tau)}_{k_{\tau,t-1}k_{\tau,t}}\right)\chi^c_{z_\tau,k_{\tau,t}}(\mathbf{x}^c_t)\chi^m_{z_\tau}(\mathbf{x}^m_t).$$

In the backward sampling step, the latent variables $\mathbf{K}$ are sequentially sampled in the reverse order as follows:

$$p(k_{\tau,d_\tau}|z_\tau, d_\tau, \mathbf{x}^c_{1:d_\tau}, \mathbf{x}^m_{1:d_\tau}) \propto \zeta_{\tau,k_{\tau,d_\tau}}, \tag{17}$$

$$p(k_{\tau,t}|z_\tau, d_\tau, k_{\tau,t+1:d_\tau}, \mathbf{x}^c_{1:d_\tau}, \mathbf{x}^m_{1:d_\tau}) \propto \zeta_{\tau,k_{\tau,t}}\phi^{(z_\tau)}_{k_{\tau,t}k_{\tau,t+1}}. \tag{18}$$

3.3.2 Sampling Model Parameters

We use the Gibbs sampling method for updating the model parameters as follows:

$$\boldsymbol{\rho} \sim \mathrm{Dirichlet}(\mathbf{a}^{\boldsymbol{\rho}} + \mathbf{b}^{\boldsymbol{\rho}}), \tag{19}$$
$$\boldsymbol{\pi}_z \sim \mathrm{Dirichlet}(\mathbf{a}^{\boldsymbol{\pi}} + \mathbf{b}^{\boldsymbol{\pi}_z}), \tag{20}$$
$$\boldsymbol{\psi} \sim \mathrm{Dirichlet}(\mathbf{a}^{\boldsymbol{\psi}} + \mathbf{b}^{\boldsymbol{\psi}}), \tag{21}$$
$$\boldsymbol{\phi}^{(z)}_k \sim \mathrm{Dirichlet}(\mathbf{a}^{\boldsymbol{\phi}} + \mathbf{b}^{\boldsymbol{\phi}^{(z)}_k}), \tag{22}$$
$$\boldsymbol{\Lambda}^c_{z,k} \sim \mathcal{W}(\mathbf{W}^c_{z,k}, \nu^c_{z,k}), \tag{23}$$
$$\boldsymbol{\mu}^c_{z,k}|\boldsymbol{\Lambda}^c_{z,k} \sim \mathcal{N}(\mathbf{m}^c_{z,k}, (\beta^c_{z,k}\boldsymbol{\Lambda}^c_{z,k})^{-1}), \tag{24}$$
$$\boldsymbol{\Lambda}^m_z \sim \mathcal{W}(\mathbf{W}^m_z, \nu^m_z), \tag{25}$$
$$\boldsymbol{\mu}^m_z|\boldsymbol{\Lambda}^m_z \sim \mathcal{N}(\mathbf{m}^m_z, (\beta^m_z\boldsymbol{\Lambda}^m_z)^{-1}), \tag{26}$$

where $\mathbf{b}^{\boldsymbol{\rho}} \in \mathbb{R}^{N_Z}$, $\mathbf{b}^{\boldsymbol{\pi}_z} \in \mathbb{R}^{N_Z}$, $\mathbf{b}^{\boldsymbol{\psi}} \in \mathbb{R}^{N_D}$, and $\mathbf{b}^{\boldsymbol{\phi}^{(z)}_k} \in \mathbb{R}^{N_K}$ are vectors counting the sampled data. $b^{\boldsymbol{\rho}}_z$ is 1 if $z = z_1$ and 0 otherwise, $b^{\boldsymbol{\pi}_z}_{z'}$ counts the number of transitions from state z to state z', $b^{\boldsymbol{\psi}}_d$ counts the number of times that sampled sections have a duration of d, and $b^{\boldsymbol{\phi}^{(z)}_k}_{k'}$ counts the number of transitions from state k to state k' in the lower-level model of section class z. The parameters $\mathbf{m}^c_{z,k}$, $\beta^c_{z,k}$, $\mathbf{W}^c_{z,k}$, and $\nu^c_{z,k}$ are calculated as follows:

$$\beta^c_{z,k} = \beta^c_0 + N_{z,k}, \quad \nu^c_{z,k} = \nu^c_0 + N_{z,k}, \tag{27}$$

$$\mathbf{m}^c_{z,k} = \frac{1}{\beta^c_{z,k}}(\beta^c_0\mathbf{m}^c_0 + N_{z,k}\overline{\mathbf{x}}^c_{z,k}), \tag{28}$$

$$(\mathbf{W}^c_{z,k})^{-1} = (\mathbf{W}^c_0)^{-1} + N_{z,k}\mathbf{S}^c_{z,k} + \frac{\beta^c_0 N_{z,k}}{\beta^c_0 + N_{z,k}}(\overline{\mathbf{x}}^c_{z,k} - \mathbf{m}^c_0)(\overline{\mathbf{x}}^c_{z,k} - \mathbf{m}^c_0)^{\mathrm{T}}, \tag{29}$$

where we have defined

$$N_{z,k} = \sum_{b=1}^{B} \delta_{z_b z}\delta_{k_b k}, \tag{30}$$

$$\overline{\mathbf{x}}^c_{z,k} = \frac{1}{N_{z,k}}\sum_{b=1}^{B} \delta_{z_b z}\delta_{k_b k}\mathbf{x}^c_b, \tag{31}$$

$$\mathbf{S}^c_{z,k} = \frac{1}{N_{z,k}}\sum_{b=1}^{B} \delta_{z_b z}\delta_{k_b k}(\mathbf{x}^c_b - \overline{\mathbf{x}}^c_{z,k})(\mathbf{x}^c_b - \overline{\mathbf{x}}^c_{z,k})^{\mathrm{T}}. \tag{32}$$

The parameters $\mathbf{m}^m_z$, β^m_z, $\mathbf{W}^m_z$, and ν^m_z can be calculated similarly.

3.3.3 Refinements

We introduce two refinements to facilitate the learning process. First, since the samples from the Gibbs sampler are not necessarily local optimums of the posterior distribution, we apply Viterbi training in the last step of the parameter estimation. Specifically, we apply the Viterbi algorithm (instead of the forward filtering-backward sampling algorithm) to estimate the latent variables and update the model parameters to the expectation values of the posterior probabilities (instead of samples from those probabilities). It is known that Viterbi training is generally efficient for finding an approximate local minimum [24].

Second, we introduce a weighting factor $w_{\mathrm{dur}}(\geq 1)$ for the duration probability to enhance its effect. Specifically, we replace the probability factor ψ_{d_b} in the forward algorithm (11) and (12) with $(\psi_{d_b})^{w_{\mathrm{dur}}}$. Similar replacements are applied to the Viterbi training step as well as to the final estimation step of latent states explained in Section 3.4. Increasing the weighting factor has the effect of lowering the temperature and putting more focus on more frequent section durations.

3.4 Estimation of Musical Sections

After training the model parameters, we compute the maximum a posteriori (MAP) estimate of the latent variables (musical sections). Specifically, we maximize the posterior probability $p(\mathbf{Z}, \mathbf{D}|\mathbf{\Theta}, \mathbf{X}^c, \mathbf{X}^m)$ with respect to latent variables $\mathbf{Z}$ and $\mathbf{D}$. This can be solved by integrating out the lower-level states $\mathbf{K}$ and applying the Viterbi algorithm for hidden semi-Markov models [25] to the upper-level model.

4. EVALUATION

4.1 Experimental Conditions

We used the RWC popular music database [26] with structure annotations [27] for evaluation. Out of the 100 pieces in the data, we used 85 musical pieces in consistent 4/4 time for simplicity. For running the proposed method, we obtained chroma features using the deep feature extractor [28] and MFCCs using the librosa library [29]. Beat information was obtained using the madmom library [23]. The empirical distribution $\mathbf{a}^{\boldsymbol{\psi}}_{\mathrm{emp}}$ of section durations was trained using the piece-wise cross validation among the 85 pieces. For parameter estimation, we iterated the Gibbs sampling 15 times and the Viterbi training 3 times, which took roughly around 5 times the duration of an input signal with a standard CPU.

The hyperparameters of the proposed models were set as follows: $N_Z = 12$, $N_D = 40$, $N_K = 16$, $\sigma = 1$, $w_{\mathrm{dur}} = 4$, $\mathbf{a}^{\boldsymbol{\rho}} = 0.1 \cdot \mathbb{I}$, $\mathbf{a}^{\boldsymbol{\pi}} = \mathbb{I}$, $\mathbf{a}^{\boldsymbol{\psi}} = 50 \cdot \mathbf{a}^{\boldsymbol{\psi}}_{\mathrm{emp}}$, $\mathbf{a}^{\boldsymbol{\phi}} = \mathbb{I}$, $\mathbf{m}^c_0 = \mathbb{E}[\mathbf{X}^c]$, $\beta^c_0 = 8$, $\mathbf{W}^c_0 = (\nu^c_0 \operatorname{cov}[\mathbf{X}^c])^{-1}$ with $\nu^c_0 = 96$, $\mathbf{m}^m_0 = \mathbb{E}[\mathbf{X}^m]$, $\beta^m_0 = 4$, and $\mathbf{W}^m_0 = (\nu^m_0 \operatorname{cov}[\mathbf{X}^m])^{-1}$ with $\nu^m_0 = 80$, where $\mathbb{I}$ denotes a vector with all entries equal to 1. The first three parameters N_Z, N_D, and N_K were determined by consulting the statistics of the annotated data (see Fig. 4). According to the data, most songs have 12 or less sections and most sections have a length of 40 beats or less. If we expect a section length of 32 beats (8 measures) and a chord duration of 2 beats, the expected number of chords in each section is 16. The value of σ is set to 1 for simplicity. The other parameters were determined by a coarse optimization w.r.t. the two evaluation measures explained below. Each parameter was optimized by a grid search, fixing the other parameters. Further optimization of the parameters is left for future work.

Method	$F_{0.5}$ (%) (segmentation)	F_{pair} (%) (clustering)
VMO [30]	8.72	28.5
CNMF [15]	17.4	41.7
SCluster [16]	23.4	45.5
Proposed	**33.0**	**54.3**

Table 1. Evaluation results.

For comparison, we ran variable Markov oracle (VMO) [30], convex non-negative matrix factorization (CNMF) [15], and spectral clustering (SCluster) [16], which were available in the music structure analysis framework (MSAF) [31]. We used the default settings in the MSAF.

We evaluated the quality of segmentation and clustering in the same way as MIREX [32]. The quality of segmentation is evaluated by the F-measure $F_{0.5}$ of section boundaries [33] defined as follows. An estimated boundary is accepted as correct if there is a boundary in the ground truth data within the range of ± 0.5 seconds. The precision rate is the percentage of correct estimates. The recall rate is the percentage of true boundaries that are correctly estimated. The F-measure $F_{0.5}$ is defined as the harmonic mean of the precision and recall.

The quality of clustering is evaluated by the pairwise F-measure F_{pair} [34] defined as follows. We compare pairs of frames (with a length of 100 ms) that are labeled with the same class in an estimation result with those in the ground truth. The precision, recall, and F-measure are defined as

$$P_{\mathrm{pair}} = \frac{|P_E \cap P_A|}{|P_E|}, \quad R_{\mathrm{pair}} = \frac{|P_E \cap P_A|}{|P_A|}, \tag{33}$$

$$F_{\mathrm{pair}} = \frac{2 P_{\mathrm{pair}} R_{\mathrm{pair}}}{P_{\mathrm{pair}} + R_{\mathrm{pair}}}, \tag{34}$$

where P_E denotes the set of similarly labeled frame pairs in the estimation and P_A denotes that in the ground truth. These values are calculated by the `mir_eval` library [35].

4.2 Experimental Results

The results in Table 1 show that SCluster had the best $F_{0.5}$ and F_{pair} among the three conventional methods. The F-measures obtained by VMO were very low and the estimated results included many unnatural short segments (see Fig. 4). This was presumably caused by the implementation in the MASF. In both $F_{0.5}$ and F_{pair}, the proposed method significantly outperformed the three compared methods.

Next, let us examine the estimated results more closely (Fig. 4). The distribution of section durations for the proposed model was similar to that of the ground truth. In particular, both distributions have peaks at the 32 beats (eight measures) and 16 beats (four measures). In contrast, the distributions for the results of the other three methods were

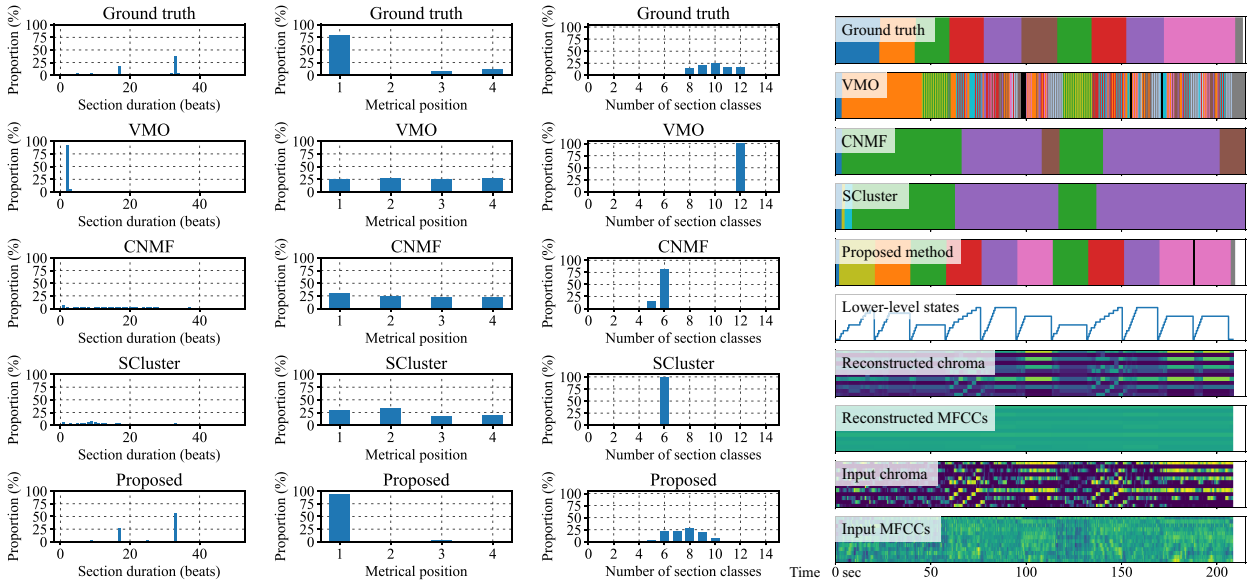

Figure 4. The left panels show the distributions of section durations, those of metrical positions of section boundaries, and those of the numbers of section classes in the estimated results and ground truth data. The right figure shows example results by the proposed and the three existing methods (RWC-MDB-P-2001 No. 29). The lower-level states are obtained by the Viterbi algorithm and the reconstructed features indicate the mean values of the corresponding output probabilities.

significantly different from that of the ground truth. This result clearly demonstrates the effect of explicitly modelling the section durations to capture their regularity. We also found that the distribution of metrical positions of section boundaries for the ground truth data was similar to that for the proposed method, but significantly different from those for the conventional methods.

The numbers of section classes in the ground truth data were roughly distributed in the range from eight to twelve. The distribution for the proposed model had a similar shape, though it is slightly shifted to the lower side. This result demonstrates the nontrivial ability of the proposed method to automatically find the appropriate number of section classes, even though it often finds the number smaller than the actual value. On the other hand, the distributions for the other methods were much more sparse; they found more or less the same number of section classes for all the tested pieces. In particular, CNMF and SCluster estimated too few section classes.

From these analyses, we find that the results of music structure analysis by the proposed method have much more similarity with the human annotated results than the compared existing methods. It is also important to point out that these results could not be made clear only by looking at the F-measures. The F-measures are not sufficient for evaluating results of music structure analysis.

We can observe these tendencies in the example results (Fig. 4). Particularly, CNMF and SCluster estimated too few section classes and irregular section durations. For the proposed method, we see that sections of the same class had similar lower-level sequences of latent states. This suggests that the model successfully captured repeated chord progressions in the sections of the same class. We can also observe that the proposed model often used only a part of lower-level states, which might be improved by imposing more constraints on the lower-level Markov chain.

For a fair comparison, we remark that parameters of the three existing methods were not optimized using our training data. Since we used limited data containing only J-pop pieces, adapting the parameters of these methods to this particular musical style may improve their performance to some extent. In addition, using the state-of-the-art beat tracker [23] to obtain reliable beat information and using that as input to those three methods may also improve their accuracy. However, it is unlikely that these methods can be refined to reproduce the aforementioned statistics of sections simply by re-training the parameters.

5. CONCLUSION

We have presented a statistical method for music structure analysis based on a Bayesian HHSMM that describes intra- and inter-section structures in a unified way. Three of the most important aspects of musical sections, homogeneity, repetitiveness, and regularity are incorporated into the model. Music segmentation and section clustering are performed jointly by unsupervised Bayesian learning of the model, and musically important characteristics such as the repetitive structure and the distribution of section durations are incorporated by the Bayesian extension. The experimental results showed that the proposed method achieved segmentation and clustering accuracies significantly better than the representative existing methods.

For future work, we plan to refine the model to incorporate the aspect of novelty and to deal with more hierarchies [16] because music has a hierarchical structure, from motive and phrase to section and section group [36]. Our unsupervised learning approach is complementary to another approach based on deep discriminative models [6–8]. A promising direction is to integrate these models into a variational autoencoding framework [22].

6. ACKNOWLEDGEMENTS

This work is supported in part by JST ACCEL No. JPMJAC1602, JSPS KAKENHI Nos. 16H01744, 19K20340, and 19H04137, and the Kyoto University Foundation.

7. REFERENCES

[1] J. Paulus, M. Müller, and A. Klapuri. State of the art report: Audio-based music structure analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636, 2010.

[2] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 452–455, 2000.

[3] K. Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Applied Signal Processing*, 2007(1):159–159, 2007.

[4] J. Serrà, M. Müller, P. Grosche, and J. Arcos. Unsupervised detection of music boundaries by time series structure features. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 1613–1619, 2012.

[5] G. Peeters and V. Bisot. Improving music structure segmentation using lag-priors. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 337–342, 2014.

[6] K. Ullrich, J. Schlüter, and T. Grill. Boundary detection in music structure analysis using convolutional neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 417–422, 2014.

[7] T. Grill and J. Schlüter. Music boundary detection using neural networks on combined features and two-level annotations. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 531–537, 2015.

[8] A. Maezawa. Music boundary detection based on a hybrid deep model of novelty, homogeneity, repetition and duration. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 206–210, 2019.

[9] G. Sargent, F. Bimbot, and E. Vincent. Estimating the structural segmentation of popular music pieces under regularity constraints. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 25(2):344–358, 2017.

[10] F. Kaiser and G. Peeters. A simple fusion method of state and sequence segmentation for music structure discovery. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 257–262, 2013.

[11] J.-J. Aucouturier and M. Sandler. Segmentation of musical signals using hidden Markov models. In *Audio Engineering Society (AES) Convention*, pages 1–8, 2001.

[12] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 127–130, 2003.

[13] M. M. Goodwin and J. Laroche. A dynamic programming approach to audio segmentation and speech/music discrimination. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 309–312, 2004.

[14] H. Grohganz, M. Clausen, N. Jiang, and M. Müller. Converting path structures into block structures using eigenvalue decompositions of self-similarity matrices. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 209–214, 2013.

[15] O. Nieto and T. Jehan. Convex non-negative matrix factorization for automatic music structure identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 236–240, 2013.

[16] B. McFee and D. P. W. Ellis. Analyzing song structure with spectral clustering. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 405–410, 2014.

[17] L. Ren, D. Dunson, S. Lindroth, and L. Carin. Dynamic nonparametric Bayesian models for analysis of music. *Journal of the American Statistical Association (JASA)*, 105(490):458–472, 2008.

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

[19] M. Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(5):1783–1794, 2006.

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

[21] T. Cheng, J. B. L. Smith, and M. Goto. Music structure boundary detection and labelling by a deconvolution of path-enhanced self-similarity matrix. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 106–110, 2018.

[22] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, pages 1–14, 2014.

[23] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. madmom: A new python audio and music signal processing library. In *ACM International Conference on Multimedia (ACMMM)*, pages 1174–1178, 2016.

[24] A. Allahverdyan and A. Galstyan. Comparative analysis of Viterbi training and maximum likelihood estimation for HMMs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1674–1682, 2011.

[25] S.-Z. Yu. Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243, 2010.

[26] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical and jazz music databases. In *International Conference on Music Information Retrieval (ISMIR)*, pages 287–288, 2002.

[27] M. Goto. AIST annotation for the RWC music database. In *International Conference on Music Information Retrieval (ISMIR)*, pages 359–360, 2006.

[28] Y. Wu and W. Li. Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 27(2):355–366, 2019.

[29] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Python in Science Conference*, pages 18–24, 2015.

[30] C.-I. Wang and G. J. Mysore. Structural segmentation with the variable Markov oracle and boundary adjustment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 291–295, 2016.

[31] O. Nieto and J. P. Bello. Systematic exploration of computational music structure research. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[32] A. F. Ehmann, M. Bay, J. S. Downie, I. Fujinaga, and D. De Roure. Music structure segmentation algorithm evaluation: Expanding on MIREX 2010 analyses and datasets. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 561–566, 2011.

[33] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *International Conference on Music Information Retrieval (ISMIR)*, pages 51–54, 2007.

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

[35] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir_eval: A transparent implementation of common MIR metrics. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[36] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1983.