MUSIC STRUCTURE ANALYSIS BASED ON AN LSTM-HSMM HYBRID MODEL

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

ABSTRACT

This paper describes a statistical music structure analysis method that splits an audio signal of popular music into musically meaningful sections at the beat level and classifies them into predefined categories such as *intro*, *verse*, and *chorus*, where beat times are assumed to be estimated in advance. A basic approach to this task is to train a recurrent neural network (*e.g.*, long short-term memory (LSTM) network) that directly predicts section labels from acoustic features. This approach, however, suffers from frequent musically unnatural label switching because the homogeneity, repetitiveness, and duration regularity of musical sections are hard to represent explicitly in the network architecture. To solve this problem, we formulate a unified hidden semi-Markov model (HSMM) that represents the generative process of homogeneous mel-frequency cepstrum coefficients, repetitive chroma features, and mel spectra from section labels, where the emission probabilities of mel spectra are computed from the posterior probabilities of section labels predicted by an LSTM. Given these acoustic features, the most likely label sequence can be estimated with Viterbi decoding. The experimental results show that the proposed LSTM-HSMM hybrid model outperformed a conventional HSMM.

1. INTRODUCTION

Music structure analysis is the fundamental task in the field of music information retrieval (MIR) [1] because the musical structure, which consists of several sections including intro, verse, bridge, and chorus, is one of the most important elements of popular music. Most studies have tackled the segmentation task, which splits audio signals into several sections [2–12], the clustering task, which categorizes such sections into several classes [13–23], or both. Beyond the clustering task that gives arbitrary labels such as "A" and "B" to detected sections, we tackle the labeling task that gives concrete labels such as "verse A", "verse B", and "chorus" [4, 24] because such musically meaningful labels are useful for playback navigation [25]. Because section labels are subjective features of music, the labeling task is still challenging. Although deep neural networks (DNNs) have widely been used for frame-level classification tasks in MIR, they often suffer from frequent musically unnatural label switching.

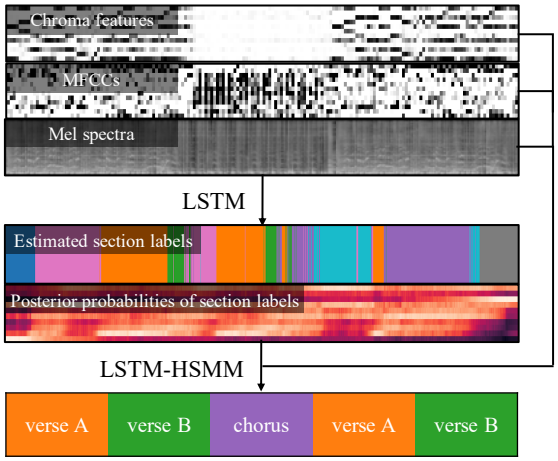

Figure 1. Proposed music structure analysis method.

In music structure analysis, the *homogeneity* and *repetitiveness* of acoustic features, the *regularity* of section durations, and the *novelty* of section boundaries, have considered as the four main noticeable aspects of musical sections [1, 11]. Using a sufficient amount of music signals with section label annotations, one could train a labeling DNN in a supervised manner such that the four aspects are learned implicitly. Another approach is to formulate a probabilistic generative model of acoustic features that can explicitly represent the four aspects and infer latent sections from observed features. A hierarchical hidden semi-Markov model (HSMM) based on the homogeneity, repetitiveness, and regularity, for example, has recently been proposed for joint segmentation and clustering [26]. The complementary properties of these approaches call for a hybrid approach for joint segmentation and labeling.

In this paper, we propose a deep generative approach to music structure analysis that integrates the labeling capability of a bidirectional long short-term memory (BLSTM) network into the classical shallow generative framework of the HSMM (Fig. 1). The unified model represents the generative processes of mel-frequency cepstrum coefficients (MFCCs) that are *homogeneous* in each section, chroma features that are *repeated* in sections of the same label, and mel spectra, from sections having *regular* durations. The

BLSTM network that estimates section labels from mel spectra at the frame level is trained in a supervised manner. The emission probabilities of mel spectra from sections are computed at run-time by referring to the posterior probabilities of section labels estimated by the network and the empirical prior distributions of section labels. Given acoustic features, the latent section sequence as well as the initial, transition, duration, and terminal probabilities of sections are estimated in a Bayesian manner with Gibbs sampling followed by Viterbi decoding, where the latent section sequence is initialized by the network to avoid bad local optima.

The main contribution of this paper is to propose a statistical joint segmentation and labeling method based on a Bayesian LSTM-HSMM hybrid model that can be adapted to each musical piece. Because the statistical characteristics of sections are specific to each musical piece, Bayesian inference based on the empirical prior distributions of those characteristics plays an essential role for improving the performance of music structure analysis. We experimentally show that the proposed method significantly outperformed a cascade model using the labeling results of the BLSTM in the post-processing and an LSTM-HSMM model using only Viterbi decoding.

2. RELATED WORK

This section reviews music structure analysis methods in terms of segmentation, clustering, and labeling.

2.1 Segmentation

In the segmentation task, the novelty plays a central role. Foote [2] detected peaks from a novelty curve obtained by convoluting a checkerboard kernel with the diagonal elements of a self-similarity matrix (SSM). Jensen [3] detected section boundaries such that a homogeneity- and novelty-aware cost function is minimized. Goto [4] and Serrà *et al.* [5] proposed novelty curves computed from lag SSMs showing repetitions as vertical lines. These methods were integrated for better segmentation [6] and the method [5] was extended for clustering [7]. Recently, Ullrich *et al.* [8] proposed a supervised method based on a convolutional neural network, which was extended to deal with coarse and fine boundary annotations [9]. Smith *et al.* [10] emphasized the importance of considering the regularity in the main analysis step, not in the post-processing step. Sargent *et al.* [11] focused on the regularity to favor comparable-size sections. Maezawa [12] used an LSTM network based on a cost function considering the homogeneity, repetitiveness, novelty, and regularity.

2.2 Clustering and Labeling

Cooper *et al.* [13] sequentially performed segmentation [2] and clustering based on intra- and inter-section characteristics. Goodwin *et al.* [14] efficiently detected off-diagonal stripes as repetitions from an SSM using dynamic programming. To deal with the repetitiveness and homogeneity, Grohganz *et al.* [15] converted a repetitiveness-aware SSM with off-diagonal stripes into a homogeneity-aware SSM with a block-diagonal structure. Nieto *et al.* [16] used a convex variant of nonnegative matrix factorization for segmentation and clustering. McFee *et al.* [17] encoded repetitive structures into a graph and performed spectral clustering for graph partitioning. Cheng *et al.* [18] converted a path-enhanced SSM into a block-enhanced SSM using nonnegative matrix factor deconvolution as in [15].

Several studies have taken a statistical approach based on generative models for joint segmentation and clustering. Aucouturier *et al.* [19] used a standard HMM. Levy *et al.* [20] proposed an HSMM based on the regularity of section durations. Ren *et al.* [21] proposed a nonparametric Bayesian HMM that can estimate an appropriate number of sections. Barrington *et al.* [22] proposed a nonparametric Bayesian switching linear dynamical system (LDS) that has the ability of automatic model complexity control.

Only a few studies have attempted to estimate musically meaningful labels. Maddage *et al.* [27] proposed a labeling method based on a typical music structure and the role of each section. Paulus *et al.* [24] performed segmentation, clustering, and labeling using a probabilistic fitness measure for the N-grams of sections.

3. PROPOSED METHOD

This section describes the proposed method for music structure analysis.

3.1 Problem Specification

The task we tackle in this paper is specified as follows:

Assumption: The beat times of a target music audio are estimated in advance by a beat tracking method [28].
Input: Beat-level chroma features $\mathbf{X}^c \triangleq \mathbf{x}^c_{1:T}$ ($\mathbf{x}^c_t \in \mathbb{R}^{12}$), MFCCs $\mathbf{X}^m \triangleq \mathbf{x}^m_{1:T}$ ($\mathbf{x}^m_t \in \mathbb{R}^{12}$), and mel spectra $\mathbf{X}^s \triangleq \mathbf{x}^s_{1:T}$ ($\mathbf{x}^s_t \in \mathbb{R}^{128}$) obtained from the target music signal, where T is the number of beats (quarter notes).
Output: Section labels $\mathbf{Z} \triangleq z_{1:N}$ ($z_n \in \{1, \ldots, K\}$) with durations $\mathbf{D} \triangleq d_{1:N}$ ($d_n \in \{1, \ldots, L\}$), where N is the number of sections, K is the number of distinct section labels, and L is the maximum number of beats in a section.

The notation $i{:}j$ represents a set of indices from i to j. Let $\mathbf{X}$ be $\{\mathbf{X}^c, \mathbf{X}^m, \mathbf{X}^s\}$, and $\mathbf{x}$ be $\{\mathbf{x}^c, \mathbf{x}^m, \mathbf{x}^s\}$.

3.2 Model Formulation

As shown in Fig. 2, we formulate a hierarchical HSMM of observed features $\mathbf{X}$ with latent sequences of section labels and abstract chord labels. Let $\mathbf{S} \triangleq \mathbf{S}_{1:N}$ be a sequence of chord sequences, where $\mathbf{S}_n \triangleq s_{n,1:d_n}$ ($s_{n,\tau} \in \{1, \ldots, M\}$) is a chord sequence in section n and M is the maximum number of chords in a section. The full probabilistic model $p(\mathbf{X}, \mathbf{Z}, \mathbf{D}, \mathbf{S})$ is defined as

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{D}, \mathbf{S}) = p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S})p(\mathbf{S}|\mathbf{Z}, \mathbf{D})p(\mathbf{Z}, \mathbf{D}), \quad (1)$$

where $p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S})$ is an acoustic model of observed features $\mathbf{X}$, $p(\mathbf{S}|\mathbf{Z}, \mathbf{D})$ is a left-to-right Markov model of chord

labels $\mathbf{S}$ and $p(\mathbf{Z}, \mathbf{D})$ is an ergodic semi-Markov model of section labels $\mathbf{Z}$ with durations $\mathbf{D}$.

3.2.1 Semi-Markov Chain of Section Labels

The ergodic semi-Markov model $p(\mathbf{Z}, \mathbf{D})$ in Eq. (1) represents the generative process of section labels $\mathbf{Z}$ and their durations $\mathbf{D}$ as follows:

$$p(\mathbf{Z}, \mathbf{D}) = p(z_1, d_1) \prod_{n=2}^{N} p(z_n, d_n | z_{n-1}, d_{n-1}), \quad (2)$$

where the individual terms are given by

$$p(z_1, d_1) = \rho_{z_1} \psi_{d_1}, \quad (3)$$

$$p(z_n, d_n | z_{n-1}, d_{n-1}) = \pi_{z_{n-1} z_n} \psi_{d_n}, \quad (4)$$

$$p(z_N, d_N | z_{N-1}, d_{N-1}) = \pi_{z_{N-1} z_N} \psi_{d_N} \upsilon_{z_N}, \quad (5)$$

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

3.2.2 Left-to-Right Markov Chain of Chord Labels

The left-to-right Markov model $p(\mathbf{S}|\mathbf{Z}, \mathbf{D})$ in Eq. (1) represents the generative process of chord labels $\mathbf{S}$ as follows:

$$p(\mathbf{S}|\mathbf{Z}, \mathbf{D}) = \prod_{n=1}^{N} p(s_{n,1}) \prod_{\tau=2}^{d_n} p(s_{n,\tau} | s_{n,\tau-1}, z_n), \quad (6)$$

where the individual terms are given by

$$p(s_{n,1} = 1) = 1, \quad (7)$$

$$p(s_{n,\tau} | s_{n,\tau-1}, z_n) = \phi^{(z_n)}_{s_{n,\tau-1} s_{n,\tau}}, \quad (8)$$

where z_n is the corresponding section label and $\phi^{(z)}_{ss'}$ is the transition probability from state s to state s'. The left-to-right Markov model meets a condition that the initial state has $s_{n,1} = 1$ and $s_{n,\tau_1} \leq s_{n,\tau_2}$ for $\tau_1 < \tau_2$. We introduce a hyperparameter σ that describes the maximum number of states that may be skipped in a transition; a transition from state s to state $s + \sigma$ is allowed. In this way, the model allows chord labels to be repeated with some variations in sections of the same label.

3.2.3 Emission of Acoustic Features

Given that the chroma features $\mathbf{X}^c$, the MFCCs $\mathbf{X}^m$, and the mel spectra $\mathbf{X}^s$ are conditionally and temporally independent, the acoustic model $p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S})$ in Eq. (1) is factorized as follows:

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{S}) = \prod_{t=1}^{T} \chi^c_{z_t, s_t}(\mathbf{x}^c_t) \chi^m_{z_t}(\mathbf{x}^m_t) \chi^s_{z_t}(\mathbf{x}^s_t), \quad (9)$$

where z_t and s_t are the section and chord labels at beat t, respectively, determined by the section-level latent variables $\mathbf{Z}$, $\mathbf{D}$, and $\mathbf{S}$, and $\chi^c_{z,s}$, χ^m_z, and χ^s_z are the emission probabilities of chroma features $\mathbf{x}^c$, MFCCs $\mathbf{x}^m$, and mel spectra $\mathbf{x}^s$, respectively.

The chroma features $\mathbf{x}^c \in \mathbb{R}^{12}$ are generated depending on both the section label z and the chord label s having the left-to-right property. The chord/chroma repetitiveness is thus represented by applying the same set of emission probabilities to all sections of the same label. The emission probability $\chi^c_{z,s}(\mathbf{x}^c)$ in Eq. (9) is given by a multivariate Gaussian distribution as follows:

$$\chi^c_{z,s}(\mathbf{x}^c) = \mathcal{N}(\mathbf{x}^c | \boldsymbol{\mu}^c_{z,s}, (\boldsymbol{\Lambda}^c_{z,s})^{-1}), \quad (10)$$

where $\boldsymbol{\mu}^c_{z,s}$ and $\boldsymbol{\Lambda}^c_{z,s}$ are a mean vector and a precision matrix, respectively.

The MFCCs $\mathbf{x}^m \in \mathbb{R}^{12}$ are generated depending on the section label z. This allows the model to capture the homogeneity of the timbral characteristics of each section. The emission probability $\chi^m_z(\mathbf{x}^m)$ in Eq. (9) is also given by a multivariate Gaussian distribution as follows:

$$\chi^m_z(\mathbf{x}^m) = \mathcal{N}(\mathbf{x}^m | \boldsymbol{\mu}^m_z, (\boldsymbol{\Lambda}^m_z)^{-1}), \quad (11)$$

where $\boldsymbol{\mu}^m_z$ and $\boldsymbol{\Lambda}^m_z$ are a mean vector and a precision matrix, respectively.

The mel spectra $\mathbf{x}^s \in \mathbb{R}^{128}$ are generated depending on the section label z. The emission probability $\chi^s_z(\mathbf{x}^s)$ in Eq. (9) is computed as follows:

$$\chi^s_z(\mathbf{x}^s) = p(\mathbf{x}^s | z) \propto \frac{p(z|\mathbf{x}^s)}{p(z)}, \quad (12)$$

where $p(z)$ is a unigram probability of section labels, and the probability $p(z|\mathbf{x}^s)$ is estimated by a labeling network (BLSTM) that infers section labels from mel spectra at the frame level. Let $p(z|\mathbf{x}^s)$ be the average of the frame-level outputs of the network in beat units.

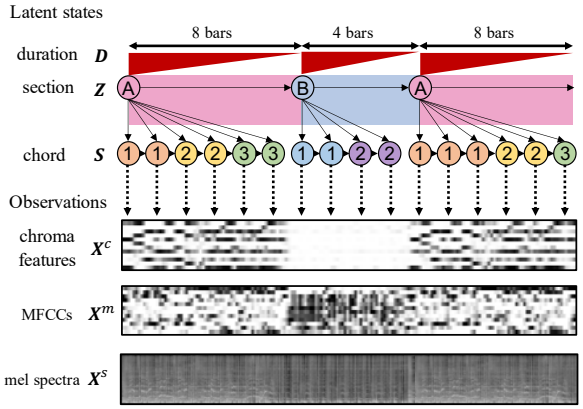

Figure 2. Proposed LSTM-HSMM hybrid model.

3.2.4 Prior Distributions Based on Musical Knowledge

To use prior knowledge about musical sections, we formulate a Bayesian HSMM by putting conjugate prior distributions on the model parameters $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\rho}, \boldsymbol{\psi}, \boldsymbol{\pi}, \boldsymbol{\upsilon}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ [26]. We put Gaussian-Wishart prior distributions on the multivariate Gaussian parameters as follows:

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

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

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. We then put Dirichlet prior distributions on the categorical parameters as follows:

$$\boldsymbol{\rho} \triangleq \rho_{1:K} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\rho}}), \quad (13)$$

$$\boldsymbol{\psi} \triangleq \psi_{1:L} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\psi}}), \quad (14)$$

$$\boldsymbol{\pi}_z \triangleq \pi_{z(1:K)} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\pi}_z}), \quad (15)$$

$$\boldsymbol{\upsilon} \triangleq \upsilon_{1:K} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\upsilon}}), \tag{16}$$

$$\boldsymbol{\phi}_s^{(z)} \triangleq \phi_{s(1:M)}^{(z)} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\phi}}), \tag{17}$$

where $\mathbf{a}^{\boldsymbol{\rho}}$, $\mathbf{a}^{\boldsymbol{\psi}}$, $\mathbf{a}^{\boldsymbol{\pi}_z}$, $\mathbf{a}^{\boldsymbol{\upsilon}}$, and $\mathbf{a}^{\boldsymbol{\phi}}$ are hyperparameters. The key advantage of Bayesian inference is that unnecessary sections can be automatically removed by controlling these sparseness-related hyperparameters.

Because "verse B" tends to come after "verse A", and section durations tend to be the integer multiples of the four measures in popular music, such a statistical tendency can be incorporated in the prior distribution. Specifically, we set $\mathbf{a}^{\boldsymbol{\rho}}$ to the empirical initial section probabilities $\mathbf{a}^{\boldsymbol{\rho}}_{\text{emp}}$, $\mathbf{a}^{\boldsymbol{\pi}_z}$ to the empirical section transition probabilities $\mathbf{a}^{\boldsymbol{\pi}_z}_{\text{emp}}$, $\mathbf{a}^{\boldsymbol{\upsilon}}$ to the empirical terminal section probabilities $\mathbf{a}^{\boldsymbol{\upsilon}}_{\text{emp}}$, and $\mathbf{a}^{\boldsymbol{\psi}}$ to the empirical section duration probabilities $\mathbf{a}^{\boldsymbol{\psi}}_{\text{emp}}$. These probabilities are multiplied by a constant factor.

3.3 Bayesian Inference

Because the posterior distribution $p(\mathbf{Z}, \mathbf{D}, \mathbf{S}, \boldsymbol{\Theta}|\mathbf{X})$ is analytically intractable, we use the Gibbs sampling method. We first sample the latent variables $\mathbf{Z}$, $\mathbf{D}$, and $\mathbf{S}$ from the distribution $p(\mathbf{Z}, \mathbf{D}, \mathbf{S}|\boldsymbol{\Theta}, \mathbf{X})$ and then sample the model parameters $\boldsymbol{\Theta}$ from the distribution $p(\boldsymbol{\Theta}|\mathbf{Z}, \mathbf{D}, \mathbf{S}, \mathbf{X})$.

3.3.1 Pretraining

We compute the empirical distributions $\mathbf{a}^{\boldsymbol{\rho}}_{\text{emp}}$, $\mathbf{a}^{\boldsymbol{\psi}}_{\text{emp}}$, $\mathbf{a}^{\boldsymbol{\pi}_z}_{\text{emp}}$, and $\mathbf{a}^{\boldsymbol{\upsilon}}_{\text{emp}}$ from training data. $\left(a^{\boldsymbol{\rho}}_{\text{emp}}\right)_z$ is the number of times that the sequence of section labels starts from a section z. $\left(a^{\boldsymbol{\psi}}_{\text{emp}}\right)_d$ is the number of times that section labels have a duration d. $\left(a^{\boldsymbol{\pi}_z}_{\text{emp}}\right)_{z'}$ is the number of transitions from a state z to a state z'. $\left(a^{\boldsymbol{\upsilon}}_{\text{emp}}\right)_z$ is the number of times that the sequence of section labels ends with a section z.

3.3.2 Initialization of Latent Variables

To avoid bad local optima, we initialize the section labels $\mathbf{Z}$ and durations $\mathbf{D}$ with the labeling network. The frame-level posterior probabilities of section labels estimated by the network are averaged in each beat t. A section label of each beat is estimated to be one that achieves the maximum value of the posterior probability at the beat. Consecutive section labels are considered as a single section; however, when the duration length of an integrated section is shorter than four beats (one bar), the section is further merged into a left or right section depending on the posterior probabilities. Because it is inefficient to deal with sections that are too long, we divide sections with a duration length longer than 32 beats into 32-beat units. After that, we perform the sampling of chord sequences $\mathbf{S}$ and model parameters $\boldsymbol{\Theta}$.

3.3.3 Sampling Latent Variables

We use the forward filtering-backward sampling algorithm for sampling $\mathbf{Z}$, $\mathbf{D}$, and $\mathbf{S}$. We introduce variables z_t and d_t that denote the section label and duration starting at beat $t-d_t+1$ and ending at beat t. We also define the marginalized emission probability for this section $\omega_{z_t}(\mathbf{x}_{t-d_t+1:t})$, which can be calculated by the forward algorithm for the Markov model of chord labels.

In the forward filtering step for the Markov model of section labels, we initialize and update the forward variables $\alpha_t(z_t, d_t) = p(z_t, d_t, \mathbf{x}_{1:t})$ as follows:

$$\alpha_t(z_t, d_t = t) = \rho_{z_t}\psi_{d_t}\omega_{z_t}(\mathbf{x}_{1:t}), \tag{18}$$

$$\alpha_t(z_t, d_t) = \sum_{z',d'} \alpha_{t-d_t}(z', d')\pi_{z'z_t}\psi_{d_t}\omega_{z_t}(\mathbf{x}_{t-d_t+1:t}). \tag{19}$$

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

$$p(z_T, d_T|\mathbf{X}) \propto \alpha_T(z_T, d_T). \tag{20}$$

When variables z_t and d_t are already sampled, the variables $z_{t'}$ and $d_{t'}$ at beat $t' = t - d_t$ are sampled according to the probability

$$p(z_{t'}, d_{t'}|z_{t:T}, d_{t:T}, \mathbf{X}) \propto \alpha_{t'}(z_{t'}, d_{t'})\pi_{z_{t'}z_t}. \tag{21}$$

Next, the chord labels $\mathbf{S}$ are sampled using the sampled $\mathbf{Z}$ and $\mathbf{D}$. Each chord sequence $\mathbf{S}_n$ is sampled by forward filtering-backward sampling for the Markov model of chord labels in section n. In the forward filtering step, we calculate the probabilities $\zeta_{n,s_{n,\tau}}$ recursively as follows:

$$\zeta_{n,s_{n,1}} = p(s_{n,1}, \mathbf{x}_1|z_n, d_n) = \delta_{s_{n,1}1}\chi_{z_n,1}(\mathbf{x}_1), \tag{22}$$

$$\zeta_{n,s_{n,\tau}} = p(s_{n,\tau}, \mathbf{x}_{1:\tau}|z_n, d_n) = \left(\sum_{s_{n,\tau-1}} \zeta_{n,s_{n,\tau-1}}\phi^{(z_n)}_{s_{n,\tau-1}s_{n,\tau}}\right)\chi_{z_n,s_{n,\tau}}(\mathbf{x}_\tau), \tag{23}$$

where $\mathbf{x}_\tau$ is a vector of observed features at the beat $\tau \in \{1, \ldots, d_n\}$ considered in relation to the section boundary, and $\chi_{z,s}(\mathbf{x})$ is a merged emission probability $p(\mathbf{x}|z, s)$. In the backward sampling step, the chord sequence $\mathbf{S}_n$ is sampled in the reverse order as follows:

$$p(s_{n,d_n}|z_n, d_n, \mathbf{x}_{1:d_n}) \propto \zeta_{n,s_{n,d_n}}, \tag{24}$$

$$p(s_{n,\tau}|z_n, d_n, s_{n,\tau+1:d_n}, \mathbf{x}_{1:d_n}) \propto \zeta_{n,s_{n,\tau}}\phi^{(z_n)}_{s_{n,\tau}s_{n,\tau+1}}. \tag{25}$$

3.3.4 Sampling Model Parameters

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

$$\boldsymbol{\rho} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\rho}} + \mathbf{b}^{\boldsymbol{\rho}}), \tag{26}$$

$$\boldsymbol{\pi}_z \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\pi}_z} + \mathbf{b}^{\boldsymbol{\pi}_z}), \tag{27}$$

$$\boldsymbol{\psi} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\psi}} + \mathbf{b}^{\boldsymbol{\psi}}), \tag{28}$$

$$\boldsymbol{\upsilon} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\upsilon}} + \mathbf{b}^{\boldsymbol{\upsilon}}), \tag{29}$$

$$\boldsymbol{\phi}_s^{(z)} \sim \text{Dirichlet}(\mathbf{a}^{\boldsymbol{\phi}} + \mathbf{b}^{\boldsymbol{\phi}_s^{(z)}}), \tag{30}$$

$$\boldsymbol{\Lambda}^c_{z,s} \sim \mathcal{W}(\mathbf{W}^c_{z,s}, \nu^c_{z,s}), \tag{31}$$

$$\boldsymbol{\mu}^c_{z,s}|\boldsymbol{\Lambda}^c_{z,s} \sim \mathcal{N}(\mathbf{m}^c_{z,s}, (\beta^c_{z,s}\boldsymbol{\Lambda}^c_{z,s})^{-1}), \tag{32}$$

$$\boldsymbol{\Lambda}^m_z \sim \mathcal{W}(\mathbf{W}^m_z, \nu^m_z), \tag{33}$$

$$\boldsymbol{\mu}^m_z|\boldsymbol{\Lambda}^m_z \sim \mathcal{N}(\mathbf{m}^m_z, (\beta^m_z\boldsymbol{\Lambda}^m_z)^{-1}), \tag{34}$$

where $\mathbf{b}^{\boldsymbol{\rho}} \in \mathbb{R}^K$, $\mathbf{b}^{\boldsymbol{\pi}_z} \in \mathbb{R}^K$, $\mathbf{b}^{\boldsymbol{\psi}} \in \mathbb{R}^L$, $\mathbf{b}^{\boldsymbol{\upsilon}} \in \mathbb{R}^K$, and $\mathbf{b}^{\boldsymbol{\phi}_s^{(z)}} \in \mathbb{R}^M$ are vectors that count the sampled data. $b^{\boldsymbol{\rho}}_z$ is 1 if $z = z_1$ and 0 otherwise, $b^{\boldsymbol{\pi}_z}_{z'}$ is the number of transitions from state z to state z', $b^{\boldsymbol{\psi}}_d$ is the number of times that sampled sections have a duration of d, $b^{\boldsymbol{\upsilon}}_z$ is 1

if $z = z_N$ and 0 otherwise, and $b_{s'}^{\boldsymbol{\phi}_s^{(z)}}$ is the number of transitions from state s to state s' in the Markov model of chord labels in section z. The parameters $\mathbf{m}_{z,s}^c$, $\beta_{z,s}^c$, $\mathbf{W}_{z,s}^c$, and $\nu_{z,s}^c$ are calculated as follows:

$$\beta_{z,s}^c = \beta_0^c + N_{z,s}, \quad \nu_{z,s}^c = \nu_0^c + N_{z,s}, \tag{35}$$

$$\mathbf{m}_{z,s}^c = \frac{1}{\beta_{z,s}^c}(\beta_0^c \mathbf{m}_0^c + N_{z,s}\overline{\mathbf{x}}_{z,s}^c), \tag{36}$$

$$(\mathbf{W}_{z,s}^c)^{-1} = (\mathbf{W}_0^c)^{-1} + N_{z,s}\mathbf{U}_{z,s}^c + \frac{\beta_0^c N_{z,s}}{\beta_0^c + N_{z,s}}(\overline{\mathbf{x}}_{z,s}^c - \mathbf{m}_0^c)(\overline{\mathbf{x}}_{z,s}^c - \mathbf{m}_0^c)^{\mathrm{T}}, \tag{37}$$

where we have defined

$$N_{z,s} = \sum_{t=1}^{T} \delta_{z_t z}\delta_{s_t s}, \tag{38}$$

$$\overline{\mathbf{x}}_{z,s}^c = \frac{1}{N_{z,s}} \sum_{t=1}^{T} \delta_{z_t z}\delta_{s_t s}\mathbf{x}_t^c, \tag{39}$$

$$\mathbf{U}_{z,s}^c = \frac{1}{N_{z,s}} \sum_{t=1}^{T} \delta_{z_t z}\delta_{s_t s}(\mathbf{x}_t^c - \overline{\mathbf{x}}_{z,s}^c)(\mathbf{x}_t^c - \overline{\mathbf{x}}_{z,s}^c)^{\mathrm{T}}. \tag{40}$$

The parameters $\mathbf{m}_z^m$, β_z^m, $\mathbf{W}_z^m$, and ν_z^m can be calculated similarly.

3.3.5 Viterbi Training

Since the samples from the Gibbs sampler are not necessarily local optima 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 [29].

3.3.6 Refinements

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 in the forward algorithm (18) and (19) with $(\psi_d)^{w_{\mathrm{dur}}}$. Similar replacements are applied to the Viterbi training step and the final estimation step of the latent states explained in Section 3.4. We also introduce a weighting factor w_{label} that balances the emission probabilities for mel spectra with the other emission probabilities. We replace the emission probability $\chi_z^s(\mathbf{x}^s)$ in (9) with $(\chi_z^s(\mathbf{x}^s))^{w_{\mathrm{label}}}$.

3.4 Estimation of Musical Sections

After training the model parameters $\boldsymbol{\Theta}$, we compute the maximum a posteriori (MAP) estimate of the musical sections. Specifically, we maximize the posterior probability $p(\mathbf{Z}, \mathbf{D}|\boldsymbol{\Theta}, \mathbf{X})$ with respect to the section labels $\mathbf{Z}$ and durations $\mathbf{D}$. This can be solved by integrating out the chord labels $\mathbf{S}$ and applying the Viterbi algorithm for HSMMs [30] to the Markov model of section labels.

4. EVALUATION

Experiments were conducted to investigate the performance of the proposed method.

4.1 Experimental Conditions

To evaluate our model, we used the 100 pieces from the RWC Popular Music Database [31] with structure annotations [32] for evaluation. We extracted chroma features using the deep feature extractor [33] and MFCCs and mel spectrograms using the librosa library [34]. Beat information was obtained using the madmom library [28]. The labeling network consisted of a single-layer BLSTM with 2048×2 cells and a fully-connected layer with output dimension K. The network was trained with 10-fold cross validation, and the empirical distributions $\mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\rho}}$, $\mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\psi}}$, $\mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\pi}_z}$, and $\mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\upsilon}}$ were trained with piece-wise cross validation for the 100 pieces. For parameter estimation, we iterated the Gibbs sampling 15 times and the Viterbi training 3 times, which took around five times longer than the duration of an input signal with a standard CPU.

The hyperparameters of the proposed model were set as follows: $K = 10$, $L = 40$, $M = 16$, $\sigma = 1$, $w_{\mathrm{dur}} = 4$, $w_{\mathrm{label}} = 0.5$, $\mathbf{a}^{\boldsymbol{\rho}} = 1{\cdot}\mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\rho}}$, $\mathbf{a}^{\boldsymbol{\pi}} = 1{\cdot}\mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\pi}}$, $\mathbf{a}^{\boldsymbol{\psi}} = 64{\cdot}\mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\psi}}$, $\mathbf{a}^{\boldsymbol{\upsilon}} = 1 \cdot \mathbf{a}_{\mathrm{emp}}^{\boldsymbol{\upsilon}}$, $\mathbf{a}^{\boldsymbol{\phi}} = \mathbb{I}$, $\mathbf{m}_0^c = \mathbb{E}[\mathbf{X}^c]$, $\beta_0^c = 64$, $\mathbf{W}_0^c = (\nu_0^c \operatorname{cov}[\mathbf{X}^c])^{-1}$ with $\nu_0^c = 512$, $\mathbf{m}_0^m = \mathbb{E}[\mathbf{X}^m]$, $\beta_0^m = 2$, and $\mathbf{W}_0^m = (\nu_0^m \operatorname{cov}[\mathbf{X}^m])^{-1}$ with $\nu_0^m = 16$, where $\mathbb{I}$ denotes a vector with all entries equal to 1. The first parameter K was determined according to the number of labels used in [24], as shown in the legend in Fig. 3. The next two parameters L and M were determined by consulting the statistics of the annotated data. In the data, most sections have a length of 40 beats or less. If we assume 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 was set to 1 for simplicity. The other parameters were determined by a coarse optimization w.r.t. the 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.

We evaluated the estimation results in terms of segmentation, clustering, and labeling. The qualities of segmentation and clustering were evaluated in the same way as MIREX [35]. The quality of segmentation was evaluated by the F-measures of section boundaries denoted by $F_{0.5}$ and $F_{3.0}$ [36]. Specifically, an estimated boundary is accepted as correct if it is within $\pm 0.5/3.0$ seconds from the ground-truth boundary. The precision rate is the percentage of correct boundaries in estimated boundaries, the recall rate is the percentage of true boundaries that are correctly estimated, and the F-measures $F_{0.5}$ and $F_{3.0}$ are defined as the harmonic means of the precision and recall rates.

The quality of clustering was evaluated by the pairwise F-measure denoted by F_{pair} [37] defined as follows. We compared 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 rate, recall rate,

Method	Segmentation		Clustering	Labeling
	$F_{0.5}$ (%)	$F_{3.0}$ (%)	F_{pair} (%)	(%)
GS3 [39]	**52.3**	73.5	54.2	n/a
SUG2 [40]	25.8	**73.7**	37.3	n/a
FK2 [41]	30.0	65.7	63.4	n/a
Paulus'09 [24]	n/a	63.0	**63.7**	34.4
Cascade model	38.3	63.9	54.9	38.8
Baseline model	30.6	53.5	43.3	39.5
Proposed model	43.3	66.5	54.6	**45.3**

Table 1. Evaluation results of a comparative experiment.

and F-measure are defined as follows:

$$P_{\text{pair}} = \frac{|P_E \cap P_A|}{|P_E|}, \quad R_{\text{pair}} = \frac{|P_E \cap P_A|}{|P_A|}, \tag{41}$$

$$F_{\text{pair}} = \frac{2P_{\text{pair}}R_{\text{pair}}}{P_{\text{pair}} + R_{\text{pair}}}, \tag{42}$$

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 using the `mir_eval` library [38]. The quality of labeling was evaluated by the accuracy in frame units, as in [24]. This is calculated by comparing a label assigned to each frame in the result and the ground truth.

For comparison in the segmentation and the clustering, we refer to GS3 [39], SUG2 [40], and FK2 [41], published in MIREX. In addition, for comparison in all three viewpoints, we quoted the result of [24] and ran two models, a cascade model and a baseline model. In the cascade model, the frame-level labels obtained by the BLSTM were counted for each cluster obtained by the HSMM [26]. The most frequently occurring label in a cluster was estimated to be the label in that cluster. Although the baseline model is similar to the proposed model, it outputs neither chroma features nor MFCCs and only uses the Viterbi decoding to obtain results.

4.2 Experimental Results

Table 1 shows the evaluation results. In the labeling accuracy, the proposed method outperformed the other methods that have the labeling ability. Compared with the cascade model using the labeling results of the BLSTM in the post-processing, the proposed model had better performance in segmentation and labeling. This indicates the effectiveness of joint segmentation and labeling in the unified probabilistic framework. In addition, compared with the baseline model using the Viterbi decoding only, the proposed model achieved better performance in all metrics. This revealed the effectiveness of piece-specific Bayesian learning based on the prior distributions. In contrast, the proposed method did not always achieve the state-of-the-art performance except for labeling. It may be because the proposed method tended to yield unnatural repetitions of the same label with various lengths. In general, sections of the same label have approximately the same length. Such a constraint could be incorporated by introducing a duration probability distribution specific for each label.

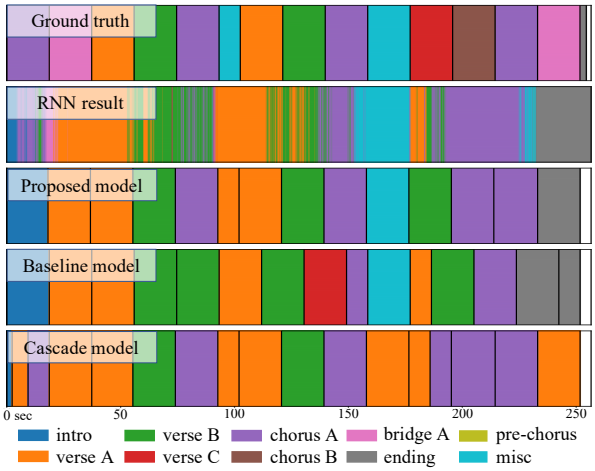

Figure 3. Example results by proposed, baseline, and cascade model (RWC-MDB-P-2001 No. 25)

Example results are shown in Fig. 3. The cascade model yielded some mislabeled sections with correctly estimated boundaries because the clustering errors of the HSMM were propagated. In contrast, because the proposed method performs segmentation and labeling simultaneously, such errors were reduced effectively. While the baseline model yielded errors originating from the errors of the BLSTM, such errors were corrected in the result of the proposed method. This suggested that the proposed method has the ability to prevent such errors by focusing on the homogeneity of MFCCs and repetitiveness of chroma features. We found that the proposed method erroneously estimated "chorus A" at the beginning of this song as "intro". Such errors could be avoided by adjusting the weights of the initial and emission probabilities or training the BLSTM with the connectionist temporal classification (CTC) loss function [42] to remove frequent musically unnatural label switching.

5. CONCLUSION

We have presented a deep generative approach to music structure analysis based on a Bayesian LSTM-HSMM hybrid model. The model represents the essential characteristics of sections, homogeneity, repetitiveness, and regularity, with MFCCs, chroma features, and mel spectra. Music segmentation and section labeling are performed jointly by unsupervised Bayesian learning of the model. The experimental results showed that the proposed method is effective for musical structure analysis.

The proposed method considers homogeneity, repetitiveness, and regularity, but not novelty, which has been emphasized in conventional research [1]. Exploiting this aspect remains an avenue for future work. It is also important to deal with further hierarchies [17], as music has a hierarchical structure moving from motives and phrases to sections and section groups [43].

6. ACKNOWLEDGEMENTS

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

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)*, 2010, pp. 625–636.

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

[3] K. Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 159–159, 2007.

[4] 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)*, vol. 14, no. 5, pp. 1783–1794, 2006.

[5] 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)*, 2012, pp. 1613–1619.

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

[7] J. Serrà, M. Müller, P. Grosche, and J. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[8] 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)*, 2014, pp. 417–422.

[9] 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)*, 2015, pp. 531–537.

[10] J. B. L. Smith and M. Goto, "Using priors to improve estimates of music structure," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 554–560.

[11] 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)*, vol. 25, no. 2, pp. 344–358, 2017.

[12] 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)*, 2019, pp. 206–210.

[13] 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)*, 2003, pp. 127–130.

[14] 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)*, 2004, pp. 309–312.

[15] 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)*, 2013, pp. 209–214.

[16] 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)*, 2013, pp. 236–240.

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

[18] 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)*, 2018, pp. 106–110.

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

[20] M. Levy and M. Sandler, "New methods in structural segmentation of musical audio," in *European Signal Processing Conference (EUSIPCO)*, 2006, pp. 1–5.

[21] L. Ren, D. Dunson, S. Lindroth, and L. Carin, "Dynamic nonparametric Bayesian models for analysis of music," *Journal of the American Statistical Association (JASA)*, vol. 105, no. 490, pp. 458–472, 2008.

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

[23] 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)*, 2013, pp. 257–262.

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

[25] ——, "Labelling the structural parts of a music piece with markov models," in *Computer Music Modeling and Retrieval (CMMR)*, 2008, pp. 166–176.

[26] G. Shibata, R. Nishikimi, E. Nakamura, and K. Yoshii, "Statistical music structure analysis based on a homogeneity-, repetitiveness-, and regularity-aware hierarchical hidden semi-markov model," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 268–275.

[27] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao, "Content-based music structure analysis with applications to music semantics understanding," in *ACM International Conference on Multimedia (ACMMM)*, 2004, pp. 112–119.

[28] 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)*, 2016, pp. 1174–1178.

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

[30] S.-Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, 2010.

[31] 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)*, 2002, pp. 287–288.

[32] M. Goto, "AIST annotation for the RWC music database," in *International Conference on Music Information Retrieval (ISMIR)*, 2006, pp. 359–360.

[33] 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)*, vol. 27, no. 2, pp. 355–366, 2019.

[34] 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*, 2015, pp. 18–24.

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

[36] 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)*, 2007, pp. 51–54.

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

[38] 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.

[39] T. Grill and J. Schlüter, "Structural segmentation with convolutional neural networks mirex submission," in *Music Information Retrieval Evaluation eXchange (MIREX)*, 2015.

[40] J. Schlüter, K. Ullrich, and T. Grill, "Structural segmentation with convolutional neural networks mirex submission," in *Music Information Retrieval Evaluation eXchange (MIREX)*, 2014.

[41] F. Kaiser and G. Peeters, "Music structural segmentation task: Ircamstructure submission," in *Music Information Retrieval Evaluation eXchange (MIREX)*, 2013.

[42] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

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