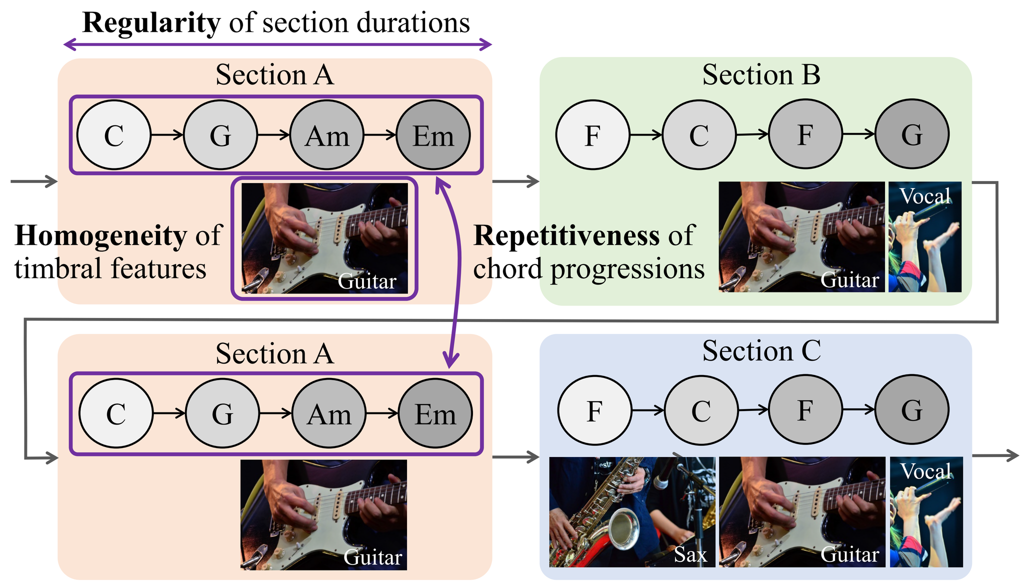
Music Structure Analysis Based on an LSTM-HSMM Hybrid Model

Go Shibata[1] Ryo Nishikimi[1] Kazuyoshi Yoshii[1]

[1] Graduate School of Informatics, Kyoto University, Japan

Abstract

- Recognize meaningful musical sections
- Probabilistic formulation based on music knowledge about musical sections: homogeneity, repetitiveness, and regularity
- Emission probabilities of mel spectra computed by bidirectional LSTM
- Unsupervised learning based on Bayesian inference

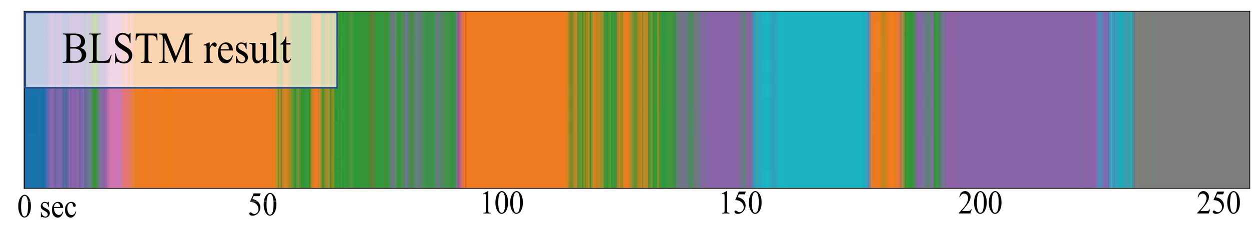


Background

Supervised learning of a DNN does not work well

- Gives unnaturally-frequent label switching because of the lack of annotated training data

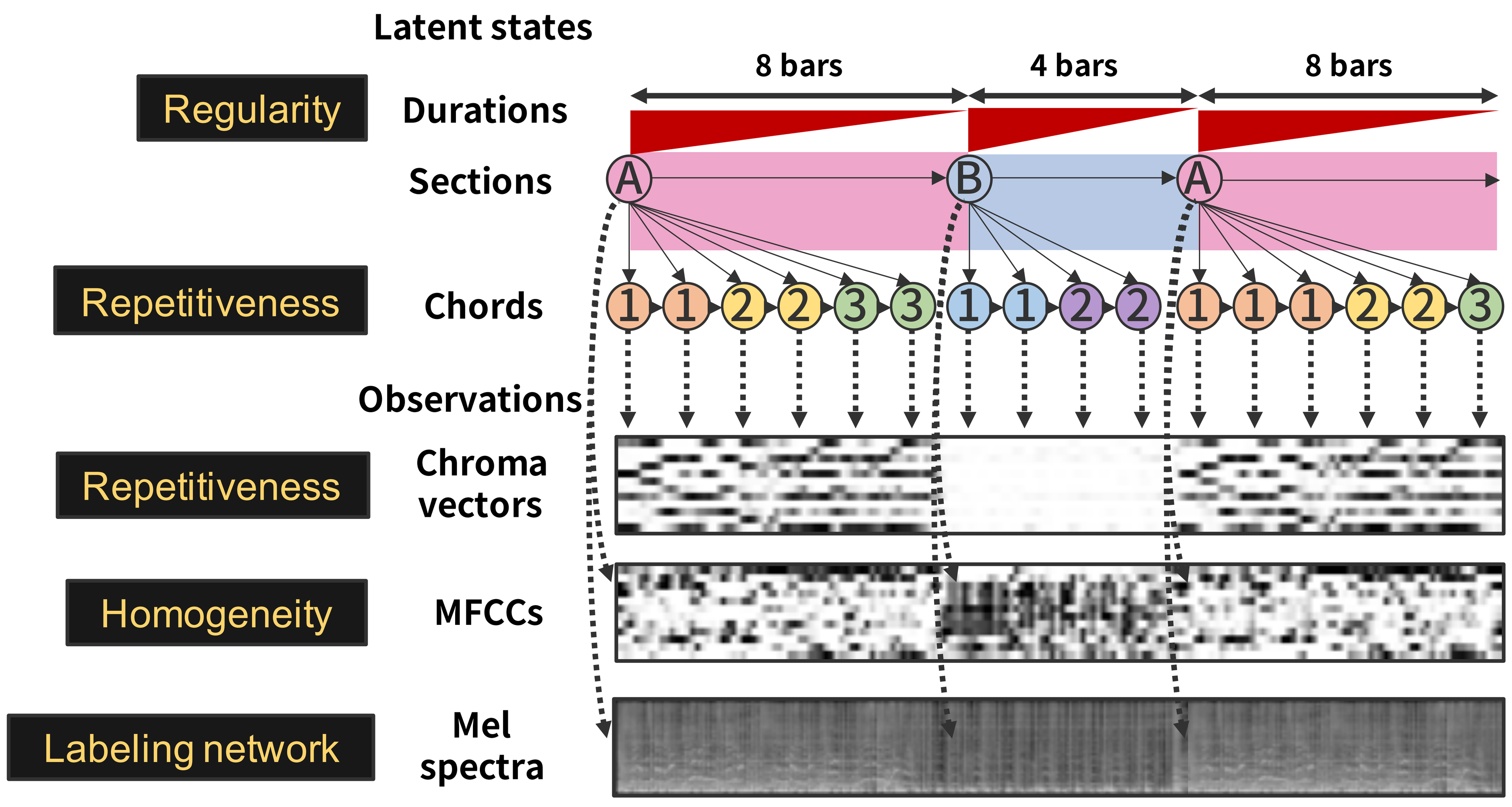
Thus we need to make effective use of music knowledge about musical sections



Model Formulation and Inference

LSTM-HSMM Hybrid Model

The proposed method deals with segmentation and labeling simultaneously



Model Formulation

Two-Level Hierarchical Markov Chains

- Upper-level: ergodic semi-Markov model

Generate a sequence of sections $\boldsymbol{Z}$ and their durations $\boldsymbol{D}$

$$p(z_1, d_1) = \rho_{z_1}\psi_{d_1}$$
$$p(z_n, d_n | z_{n-1}, d_{n-1}) = \pi_{z_{n-1}z_n}\psi_{d_n}$$
$$p(z_N, d_N | z_{N-1}, d_{N-1}) = \pi_{z_{N-1}z_N}\psi_{d_N}\upsilon_{z_N}$$

- ρ_z initial probability of sections
- ψ_d duration probability of sections
- $\pi_{zz'}$ transition probability of sections
- υ_z terminal probability of sections

- Lower-level: left-to-right Markov model

Generate a sequence of chords $\boldsymbol{S}$

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

- $\phi^{(z)}_{ss'}$ transition probability of chords
- $s_{n,1} = 1$
- $\tau_1 < \tau_2 \Longrightarrow s_{n,\tau_1} \le s_{n,\tau_2}$

Acoustic Model

Outputs chroma vectors, MFCCs, and mel spectra

- Chroma vectors

Describe the repetitiveness of chord sequences

Generated from Gaussian distributions corresponding to sections and chords

$$\boldsymbol{x}^c \sim \mathcal{N}\left(\boldsymbol{\mu}^c_{z,s}, \left(\boldsymbol{\Lambda}^c_{z,s}\right)^{-1}\right)$$

- MFCCs (Mel-Frequency Cepstrum Coefficients)

Describe the homogeneity of timbral features

Generated from Gaussian distributions corresponding to sections

$$\boldsymbol{x}^m \sim \mathcal{N}(\boldsymbol{\mu}^m_z, (\boldsymbol{\Lambda}^m_z)^{-1})$$

- Mel spectra

Associate sections with labels

Generated using probabilities based on a bidirectional LSTM

$$\boldsymbol{x}^s \sim p(z|\boldsymbol{x}^s)/p(z)$$

Prior Distributions

We put conjugate prior distributions for parameters of the model

- Degenerate unnecessary sections during the Bayesian sparse learning
- Incorporate the regularity of section durations as prior knowledge
- Use empirical distributions $\boldsymbol{\alpha}^{\rho}_{\text{emp}}, \boldsymbol{\alpha}^{\pi_z}_{\text{emp}}, \boldsymbol{\alpha}^{\psi}_{\text{emp}}$, and $\boldsymbol{\alpha}^{\upsilon}_{\text{emp}}$ for the prior distributions

Inference

Gibbs Sampling

Sampling latent variables	Sampling model parameters
Forward filtering-backward sampling	Sample from posterior distributions of parameters
$p(\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S}\mid\boldsymbol{\Theta},\boldsymbol{X}^c,\boldsymbol{X}^m,\boldsymbol{X}^s)$	$p(\boldsymbol{\Theta}\mid\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S},\boldsymbol{X}^c,\boldsymbol{X}^m,\boldsymbol{X}^s)$
Sample	Sample
$\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S}$	$\boldsymbol{\Theta} = \{\boldsymbol{\rho},\boldsymbol{\psi},\boldsymbol{\pi},\boldsymbol{\upsilon},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\Lambda}\}$

Iterate a sufficient number of times

Viterbi Training

Sampling latent variables	Sampling model parameters
Viterbi algorithm	Take expectation values of posteriors
$p(\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S}\mid\boldsymbol{\Theta},\boldsymbol{X}^c,\boldsymbol{X}^m,\boldsymbol{X}^s)$	$p(\boldsymbol{\Theta}\mid\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S},\boldsymbol{X}^c,\boldsymbol{X}^m,\boldsymbol{X}^s)$
Maximize	Expectation
$\boldsymbol{Z},\boldsymbol{D},\boldsymbol{S}$	$\boldsymbol{\Theta} = \{\boldsymbol{\rho},\boldsymbol{\psi},\boldsymbol{\pi},\boldsymbol{\upsilon},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\Lambda}\}$

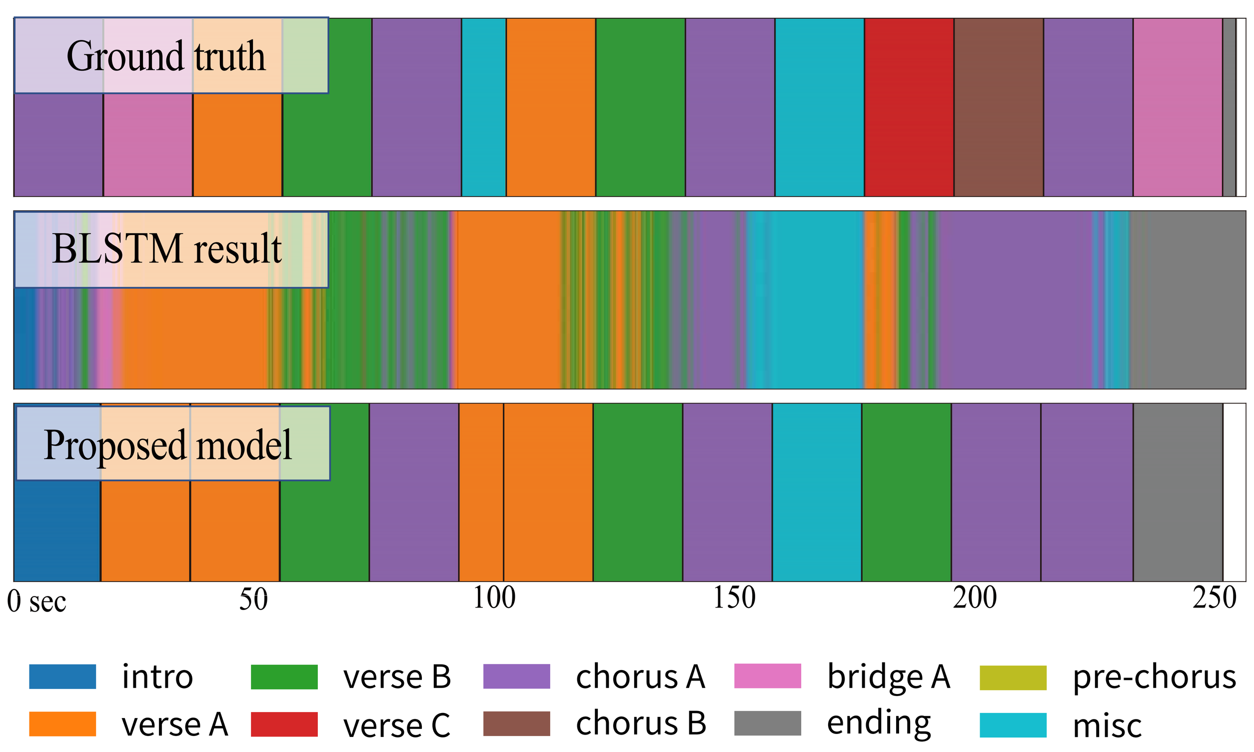
Evaluation

Comparison with Conventional Methods

Method	Segmentation		Clustering	Labeling
	$F_{0.5}$ (%)	$F_{3.0}$ (%)	F_{pair} (%)	accuracy(%)
GS3 [Grill+, '15]	**52.3**	73.5	54.2	n/a
SUG2 [Schlüter+, '14]	25.8	**73.7**	37.3	n/a
FK2 [Kaiser+, '13]	30.0	65.7	63.4	n/a
[Paulus, '09]	n/a	63.0	**63.7**	34.4
Proposed	43.3	66.5	54.6	**45.3**

- The proposed method worked best in terms of labeling accuracy
- There is much room for improvement except for labeling accuracy

Example of Analysis Result



- The bidirectional LSTM was confused between "verse A" and "verse B" or between "verse B" and "chorus A", while the proposed model correctly recognized these sections
- We need to improve the proposed model to avoid errors such as the confusion between "intro" and "chorus A"

Future Work

- Refine the model to incorporate the novelty aspect
- Deal with more hierarchies because music has a hierarchical structure, from motive and phrase to section and section group