JOINT TRANSCRIPTION OF LEAD, BASS, AND RHYTHM GUITARS BASED ON A FACTORIAL HIDDEN SEMI-MARKOV MODEL

Kentaro Shibata[1] *Ryo Nishikimi*[1] *Satoru Fukayama*[2] *Masataka Goto*[2]
Eita Nakamura[1] *Katsutoshi Itoyama*[1] *Kazuyoshi Yoshii*[1]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper describes a statistical method for estimating musical scores for lead, bass, and rhythm guitars from polyphonic audio signals of typical band-style music. To perform multi-instrument transcription involving multi-pitch detection and part assignment, it is crucial to formulate a musical language model that represents the characteristics of each part in order to solve the ambiguity of part assignment and estimate a musically-natural score. We propose a factorial hidden semi-Markov model that consists of three language models corresponding to the three guitar parts (three latent chains) and an acoustic model of a mixture spectrogram (emission model). The language model for rhythm guitar represents a *homophonic* sequence of musical notes (chord sequence) and those for lead and bass guitars represent a *monophonic* sequence of musical notes in a *higher* and *lower* frequency range respectively. The acoustic model represents a spectrogram as a sum of low-rank spectrograms of the three guitar parts approximated by NMF. Given a spectrogram, we estimate the note sequences using Gibbs sampling. We show that our model outperforms a state-of-the-art multi-pitch detection method in the accuracy and naturalness of the transcribed scores.

***Index Terms*—** Automatic music transcription, multi-pitch estimation, multi-instrument transcription, HSMM, and NMF.

1. INTRODUCTION

Transcribing a musical score from an audio signal is a fundamental and challenging problem in music information processing [1]. From a practical viewpoint, it is important to develop a transcription method for band music to facilitate widely-enjoyed cover performances. A typical band playing popular music (*e.g.*, The Beatles) consists of vocal, several guitars, and drum parts. Transcription of vocal and drum parts have each been studied extensively, and high accuracies have been reported [2–7]. We therefore focus on accompaniment parts that are typically played by three guitars (lead, bass, and rhythm guitars).

Accompaniment-part transcription for band music is a challenging task because multi-pitch detection and part assignment of each note are both required. A major approach to multi-pitch detection is to use probabilistic latent component analysis (PLCA) or non-negative matrix factorization (NMF) based on the sparseness and low-rankness of source spectrograms. For part assignment, these methods have been extended to use pre-learned spectral templates of multiple instruments [8–13]. More recently, end-to-end neural networks have been applied successfully to executing multi-pitch detection and part assignment simultaneously [14, 15].

This work was supported in part by JST ACCEL No. JPMJAC1602, JSPS KAKENHI No. 16H01744 and No. 16J05486, and the Kyoto University Foundation.

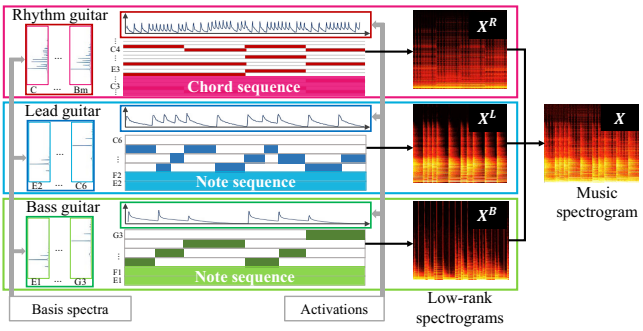

Fig. 1. A factorial hidden semi-Markov model (FHSMM) that represents the generative process of audio signals of three guitar parts of band music.

Since most of the previous methods rely on the *timbral* characteristics of instruments, they struggle with band music with multiple instruments having similar timbre (*e.g.*, two guitars). Recently, attempts have been made to incorporate a music language model representing a musical grammar (*e.g.*, sequential dependency of chords and musical notes) to complement the acoustic model in such difficult situations. Although Sigtia *et al.* [14] proposes an end-to-end polyphonic transcription method based on a recurrent neural network (RNN)-based language model, the model's effect is essentially smoothing since it is defined at the *frame level*. In order to get well-formed transcriptions, as they mentioned, a beat-level language model is considered to be effective. Schramm *et al.* [16] combines a PLCA-based acoustic model with a hidden Markov model (HMM)-based musical language model (also defined at the frame level), where multiple vocal parts were treated independently.

In this paper, we propose a transcription method for the accompaniment part of band music based on the *beat-level sequential* characteristics of accompanying instruments. We assume that three kinds of guitars—lead, bass, and rhythm—are used in band music and have different sequential characteristics, as listed in Table 1. Our method uses a factorial hidden semi-Markov model (FHSMM) [17–19] that consists of three language models (semi-Markov models), corresponding to the three guitar parts, and an NMF-based acoustic model of a mixture spectrogram (Fig. 1). The language model for the rhythm guitar represents a *homophonic* sequence of musical notes (chord sequence), while those for the lead and bass guitars represent a *monophonic* sequence of musical notes in a *higher* or *lower* frequency range respectively. The acoustic model represents a spectrogram as a sum of the spectrograms of the three guitar parts, each of which is represented as a low-rank matrix obtained by the product of an activation vector and basis spectra. A key feature of the

Table 1. The characteristics of thee guitar parts.

Part	Sequence	Frequency range	Rhythm
Lead	Monophonic	Middle - High	Complex
Bass	Monophonic	Low	Complex
Rhythm	Homophonic	Wide	Simple

acoustic model is that only one of the basis spectra is allowed to be activated in each tatum in each guitar part. Given a mixture spectrogram as observed data, the basis spectra, activations, and latent chains are statistically estimated using Gibbs sampling. The musical score for each guitar is then obtained by using the Viterbi algorithm.

A major contribution of this study is to achieve joint transcription of multiple musical instruments that have similar timbral characteristics. More specifically, we propose an integrated language and acoustic model that can describe beat-level symbolic musical grammar and dependency between multiple instrument parts. We also show that the initialization problem of the NMF-based model can be improved by a support from a DNN-based model, which has strong capability of expressing acoustic signals. We achieve an improvement on the accuracy, which is an important step towards developing a practical transcription method that can be used for popular music.

2. PROPOSED METHOD

Our method jointly performs multi-pitch estimation and part assignment in a unified framework. We formulate a generative model of a music spectrogram and the corresponding musical score and then solve the inverse problem. That is, given a music spectrogram, we estimate the score described as latent variables in the model. The problem is defined as follows:

Input: The magnitude spectrogram of a target signal $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ and 16th-note-level tatum times
Output: Musical scores of the three guitar parts

Here, F is the number of frequency bins, T is the number of time frames. In this paper we assume that the time signature of the target signal is 4/4.

2.1. Probabilistic Factorial Modeling

We represent a music spectrogram $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ as a sum of the spectrograms $\mathbf{X}^{\mathrm{R}}, \mathbf{X}^{\mathrm{L}}$, and $\mathbf{X}^{\mathrm{B}} \in \mathbb{R}_+^{F\times T}$ of the rhythm, lead, and bass guitars (Fig. 1):

$$x_{ft} = x_{ft}^{\mathrm{R}} + x_{ft}^{\mathrm{L}} + x_{ft}^{\mathrm{B}}. \tag{1}$$

The generative processes of $\mathbf{X}^{\mathrm{R}}$, $\mathbf{X}^{\mathrm{L}}$, and $\mathbf{X}^{\mathrm{B}}$ are represented by rhythm, lead, and bass guitar models, respectively. In each guitar model, the corresponding musical score is represented as a latent chain of variables, whose generative process is described by a semi-Markov model. Given the three chains, the spectrogram $\mathbf{X}^{\mathrm{R}}$, $\mathbf{X}^{\mathrm{L}}$, and $\mathbf{X}^{\mathrm{B}}$ are generated by an NMF-based acoustic model conditioned on the musical score. Finally, the observed spectrogram $\mathbf{X}$ is generated according to Eq. (1). The generative process for each guitar part is described as a hidden semi-Markov model (HSMM) as explained below. The total model is a factorial HSMM with three latent chains.

2.1.1. Language Model

We formulate a rhythm guitar HSMM, representing the generative process of $\mathbf{X}^{\mathrm{R}}$, and lead and bass guitar *conditional* HSMMs, representing the generative process of $\mathbf{X}^{\mathrm{L}}$ and $\mathbf{X}^{\mathrm{B}}$ given the chord sequence specified by rhythm guitar (Fig. 2). For convenience, "$\star$" will be used to represent "L" (lead guitar) or "B" (bass guitar) and

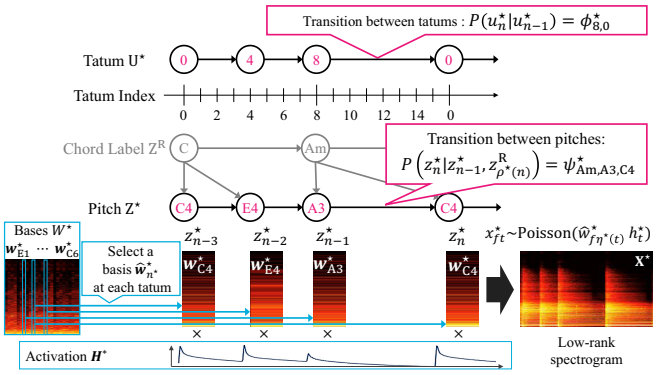

Fig. 2. The generative process of spectrograms for lead/bass guitar based on a conditional HSMM. $\star$ indicates lead (L) or bass (B).

"$\bullet$" will be used to represent "R" (rhythm guitar), "L", or "B". The latent chain of the rhythm guitar is specified by a sequence of chord symbols $\mathbf{Z}^{\mathrm{R}} = \{z_1^{\mathrm{R}}, \ldots, z_{N^{\mathrm{R}}}^{\mathrm{R}}\}$ with relative onset positions $\mathbf{U}^{\mathrm{R}} = \{u_1^{\mathrm{R}}, \ldots, u_{N^{\mathrm{R}}}^{\mathrm{R}}\}$, where N^{R} is the number of chords, z_n^{R} takes one of 24 values of $\{\mathrm{C}, \ldots, \mathrm{B}\} \times \{\mathrm{major}, \mathrm{minor}\}$, and u_n^{R} takes an integer from 0 to 15 indicating a position on the 16th-note-level tatum grid in a measure. Likewise, the latent chain of the lead or bass guitar's HSMM is specified by a sequence of pitches $\mathbf{Z}^\star = \{z_1^\star, \ldots, z_{N^\star}^\star\}$ with relative onset positions $\mathbf{U}^\star = \{u_1^\star, \ldots, u_{N^\star}^\star\}$, where $N^\star$ is the number of musical notes z_n^{L} takes one of 45 pitches of $\{\mathrm{E2}, \ldots, \mathrm{C6}\}$, z_n^{B} takes one of 28 pitches of $\{\mathrm{E1}, \ldots, \mathrm{G3}\}$, and $u_n^\star$ takes an integer from 0 to 15.

The sequence of chord symbols $\mathbf{Z}^{\mathrm{R}}$ and the sequence of pitches $\mathbf{Z}^\star$ are represented by a Markov model as follows:

$$p\left(z_1^{\mathrm{R}}|\boldsymbol{\pi}^{\mathrm{R}}\right) = \pi_{z_1^{\mathrm{R}}}^{\mathrm{R}}, \quad p\left(z_n^{\mathrm{R}}|z_{n-1}^{\mathrm{R}}, \boldsymbol{\psi}^{\mathrm{R}}\right) = \psi_{z_{n-1}^{\mathrm{R}}, z_n^{\mathrm{R}}}^{\mathrm{R}}, \tag{2}$$

$$p\left(z_1^\star|\mathbf{Z}^{\mathrm{R}}, \mathbf{U}^{\mathrm{R}}, \boldsymbol{\pi}^\star\right) = \pi_{z_1^{\mathrm{R}}, z_1^\star}^\star, \tag{3}$$

$$p\left(z_n^\star|z_{n-1}^\star, \mathbf{Z}^{\mathrm{R}}, \mathbf{U}^{\mathrm{R}}, \boldsymbol{\psi}^\star\right) = \psi_{z_{\rho^\star(n)}^{\mathrm{R}}, z_{n-1}^\star, z_n^\star}^\star, \tag{4}$$

where π_a^{R} is the initial probability of chord a, $\psi_{a,b}^{\mathrm{R}}$ is the transition probability from chord a to chord b, $\rho^\star(n)$ indicates a chord to which musical note $z_n^\star$ belongs, $\pi_{c,a}^\star$ is the initial probability of pitch a given chord c, and $\psi_{c,a,b}^\star$ is the transition probability from pitch a to pitch b given chord c. This formulation is based on the fact that the pitches of the lead and bass guitars are strongly correlated with the chords of the rhythm guitar.

The sequence of chord onset times $\mathbf{U}^{\mathrm{R}}$ and note onset times $\mathbf{U}^\star$ are represented by metrical Markov models [20, 21] as follows:

$$p(u_n^\bullet|u_{n-1}^\bullet, \boldsymbol{\phi}^\bullet) = \phi_{u_{n-1}^\bullet, u_n^\bullet}^\bullet, \tag{5}$$

where $\phi_{a,b}^\bullet$ is the transition probability from tatum position a to tatum position b. Note that if $a \geq b$, the chord or note continues over a bar line. The maximum duration of a chord or note is restricted to the measure length (16 tatums). When combined with the models for pitches/chords, this is a semi-Markov model because the time unit of state transition is different from the tatum unit and the metrical Markov model describes the distribution of sojourn times of latent states.

2.1.2. Acoustic Model

The generative process of the spectrograms of three guitars are formulated in the same way. We use the probabilistic formulation of NMF based on the Kullback-Leibler divergence (KL-NMF) [13].

The spectrogram of the rhythm, lead, or bass guitar part $\mathbf{X}^\bullet$ is generated by using a set of basis spectra $\mathbf{W}^\bullet$ and an activation vector $\mathbf{h}^\bullet \in \mathbb{R}_+^T$. The generative process of the mixture spectrogram is described as follows:

$$p(x_{ft}|\mathbf{Z},\mathbf{U},\mathbf{W},h_t) = \mathrm{Poisson}\left(x_{ft}\,\middle|\,\sum_{\bullet=\mathrm{R,L,B}} w^\bullet_{f\eta^\bullet(t)} h^\bullet_t\right), \quad (6)$$

where $\eta^\bullet(t)$ indicates the musical note to which frame t belongs and is determined by $\mathbf{Z}^\bullet$ and $\mathbf{U}^\bullet$. Basis spectra $\mathbf{W}^\bullet$ are given by $\mathbf{W}^\mathrm{R} = \{\mathbf{w}^\mathrm{R}_\mathrm{C},\ldots,\mathbf{w}^\mathrm{R}_\mathrm{Bm}\} \in \mathbb{R}_+^{F\times 24}$,$\mathbf{W}^\mathrm{L} = \{\mathbf{w}^\mathrm{L}_\mathrm{E2},\ldots,\mathbf{w}^\mathrm{L}_\mathrm{C6}\} \in \mathbb{R}_+^{F\times 33}$ or $\mathbf{W}^\mathrm{B} = \{\mathbf{w}^\mathrm{B}_\mathrm{E1},\ldots,\mathbf{w}^\mathrm{B}_\mathrm{G3}\} \in \mathbb{R}_+^{F\times 28}$, where $\mathbf{w}^\mathrm{R}_\chi$ indicates the basis spectrum of chord χ and $\mathbf{w}^\star_\chi$ indicates the basis spectrum of note χ. A key feature of our model is that only a single basis spectrum specified by $\eta^\bullet(t)$ is activated in each frame t for generating the spectrogram $\mathbf{X}^\bullet$.

2.2. Bayesian Formulation

To integrate the sub-models described in Sections 2.1.1, and 2.1.2, we formulate a semi-Bayesian model given by

$$\begin{aligned} &p(\mathbf{X},\mathbf{Y};\boldsymbol{\Theta}) \\ &= p(\mathbf{X}|\mathbf{Z},\mathbf{U},\mathbf{W},\mathbf{H})p(\mathbf{Z}|\boldsymbol{\pi},\boldsymbol{\psi})p(\mathbf{U}|\boldsymbol{\phi})p(\mathbf{W})p(\mathbf{H}), \end{aligned} \quad (7)$$

where $\boldsymbol{\Theta} = \{\pi^\mathrm{R},\pi^\mathrm{L},\pi^\mathrm{B},\psi^\mathrm{R},\psi^\mathrm{L},\psi^\mathrm{B},\phi^\mathrm{R},\phi^\mathrm{L},\phi^\mathrm{B}\}$ is parameters that are trained in advance and $\mathbf{Y} = \{\mathbf{Z},\mathbf{U},\mathbf{W},\mathbf{H}\}$ is random variables estimated for an observed spectrogram $\mathbf{X}$ during runtime. Here, $\mathbf{Z} = \{\mathbf{Z}^\mathrm{R},\mathbf{Z}^\mathrm{L},\mathbf{Z}^\mathrm{B}\}$ and $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{H}$ are defined similarly.

We introduce prior distributions $p(\mathbf{W})$ and $p(\mathbf{H})$ to make $\mathbf{W}$ close to the template bases and make $\mathbf{H}$ sparse. We use gamma priors for $\mathbf{W}$ as follows:

$$w^\bullet_{fk} \sim \mathcal{G}\left(a_{w^\bullet_{fk}}, b_{w^\bullet_{fk}}\right), \quad (8)$$

where $k \in \{\mathrm{C},\ldots,\mathrm{Bm}\}$, $\{\mathrm{E2},\ldots,\mathrm{C6}\}$, or $\{\mathrm{E1},\ldots,\mathrm{G3}\}$ denotes a chord or a pitch and a_w and b_w are the shape and rate hyperparameters. We use weak priors on $\mathbf{W}^\mathrm{R}$ (a^R_w and b^R_w are smaller than $a^\star_w$ and $b^\star_w$) to handle the large variance of the rhythm guitar. The hyperparameters are set in a way that the prior expectation of $\mathbf{W}$ matches the template spectra of chords and musical notes prepared in advance. Similarly, we use gamma priors for $\mathbf{H}$ as follows:

$$h^\bullet_t \sim \mathcal{G}\left(a_{h^\bullet_t}, b_{h^\bullet_t}\right), \quad (9)$$

where $a^\bullet_{h_t}$ and $b^\bullet_{h_t}$ are hyperparameters.

2.3. Bayesian Inference

Given a music spectrogram $\mathbf{X}$, we aim to calculate the posterior distribution according to Bayes' theorem:

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\Theta}) = p(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta})/p(\mathbf{X}|\boldsymbol{\Theta}). \quad (10)$$

Since this distribution is analytically intractable, we use Gibbs sampling for alternately and iteratively sampling the latent variables $\mathbf{Z}$ and $\mathbf{U}$, the basis spectra $\mathbf{W}$, and the activations $\mathbf{H}$. That is, we get samples of $\mathbf{G} \subset \mathbf{Y}$ from a conditional posterior distribution $p(\mathbf{G}|\mathbf{Y}_{\neg\mathbf{G}},\mathbf{X},\boldsymbol{\Theta})$, where $\mathbf{Y}_{\neg\mathbf{G}}$ indicates the subset of $\mathbf{Y}$ obtained by removing $\mathbf{G}$ from $\mathbf{Y}$. We henceforth do not write the dependency on $\boldsymbol{\Theta}$ for brevity.

2.3.1. Updating Latent Variables $\mathbf{Z}$ *and* $\mathbf{U}$

To sample $\mathbf{Z}^\mathrm{R}$ and $\mathbf{U}^\mathrm{R}$ from $p(\mathbf{Z}^\mathrm{R},\mathbf{U}^\mathrm{R}|\mathbf{Y}_{\neg\mathbf{Z}^\mathrm{R},\mathbf{U}^\mathrm{R}},\mathbf{X})$, an efficient forward filtering-backward sampling algorithm can be used. In forward filtering, a forward message of the rhythm guitar part $\alpha^\mathrm{R}(z^\mathrm{R}_n,u^\mathrm{R}_n)$ is calculated recursively as follows:

$$\alpha^\mathrm{R}\left(z^\mathrm{R}_1,u^\mathrm{R}_1\right) = p\left(z^\mathrm{R}_1\right) = \pi^\mathrm{R}_{z^\mathrm{R}_1}, \quad (11)$$

$$\begin{aligned} \alpha^\mathrm{R}\left(z^\mathrm{R}_n,u^\mathrm{R}_n\right) &= p\left(x_{\tau^\mathrm{R}(n-1):\tau^\mathrm{R}(n)+1}|z^\mathrm{R}_n,u^\mathrm{R}_n\right) \\ &\cdot \sum_{z^\mathrm{R}_{n-1}}\sum_{u^\mathrm{R}_{n-1}} \psi^\mathrm{R}_{z^\mathrm{R}_n,z^\mathrm{R}_{n-1}}\phi^\mathrm{R}_{u^\mathrm{R}_n,u^\mathrm{R}_{n-1}}\alpha\left(z^\mathrm{R}_{n-1},u^\mathrm{R}_{n-1}\right), \end{aligned} \quad (12)$$

where $\tau^\mathrm{R}(n)$ is the last frame of the n-th chord and $x_{a:b}$ is $\{x_a,\ldots,x_b\}$.

In the backward sampling step, z^R_n and u^R_n are sampled recursively by calculating a backward message $\gamma^\mathrm{R}(z^\mathrm{R}_n,u^\mathrm{R}_n)$ as follows:

$$\gamma^\mathrm{R}\left(z^\mathrm{R}_N,u^\mathrm{R}_N\right) = p\left(z^\mathrm{R}_N,u^\mathrm{R}_N|\mathbf{X}\right) \propto \alpha^\mathrm{R}\left(z^\mathrm{R}_N,u^\mathrm{R}_N\right), \quad (13)$$

$$\begin{aligned} \gamma^\mathrm{R}\left(z^\mathrm{R}_n,u^\mathrm{R}_n\right) &= p\left(z^\mathrm{R}_n,u^\mathrm{R}_n|z^\mathrm{R}_{n+1:N},u^\mathrm{R}_{n+1:N},\mathbf{X}\right) \\ &\propto \psi^\mathrm{R}_{z^\mathrm{R}_n,z^\mathrm{R}_{n+1}}\phi^\mathrm{R}_{u^\mathrm{R}_n,u^\mathrm{R}_{n+1}}\alpha\left(z^\mathrm{R}_n,u^\mathrm{R}_n\right). \end{aligned} \quad (14)$$

Latent variables $\mathbf{Z}^\star$ and $\mathbf{U}^\star$ are updated similarly except that the transition probability $\psi^\mathrm{R}_{a,b}$ is replaced with $\psi^\star_{c,a,b}$.

2.3.2. Updating Basis Spectra $\mathbf{W}$ *and Activations* $\mathbf{H}$

Similarly as the Bayesian inference of NMF [22], $\mathbf{W}$ and $\mathbf{H}$ are sampled directly from the conditional posterior distribution $p(\mathbf{W},\mathbf{H}|\mathbf{Z},\mathbf{U},\mathbf{X})$. Let $\lambda^\bullet_{ft}$ be an auxiliary variable, calculated from the latest samples of $\mathbf{W}$ and $\mathbf{H}$ as

$$\lambda^\bullet_{ft} = \frac{w^\bullet_{f\eta(t)}h^\bullet_t}{w^\mathrm{R}_{f\eta^\mathrm{R}(t)}h^\mathrm{R}_t + w^\mathrm{L}_{f\eta^\mathrm{L}(t)}h^\mathrm{L}_t + w^\mathrm{B}_{f\eta^\mathrm{B}(t)}h^\mathrm{B}_t}. \quad (15)$$

Using $\boldsymbol{\lambda}$, $\mathbf{W}$ and $\mathbf{H}$ are sampled as follows:

$$w^\bullet_{fk} \sim \mathcal{G}\left(a^\bullet_w + \sum_t x_{ft}\lambda^\bullet_{ft},\ \ b^\bullet_w + \sum_t h^\bullet_t\right), \quad (16)$$

$$h^\bullet_t \sim \mathcal{G}\left(a^\bullet_{h^\bullet_t} + \sum_f x_{ft}\lambda^\bullet_{ft},\ \ b^\bullet_{h^\bullet_t} + \sum_f w^\bullet_{fk}\right). \quad (17)$$

2.4. Musical Score Estimation

We obtain the most likely $\mathbf{Z}^\bullet$ and $\mathbf{U}^\bullet$ by the Viterbi algorithm and the final estimate of $\mathbf{H}$ by taking the mean of the posterior distribution in Eq. (17). Musical scores for the three guitar parts are estimated with these variables. The musical score for the lead or bass guitar can be obtained from the pitches $\mathbf{Z}^\star$ and onset times $\mathbf{U}^\star$. Note that a series of musical notes with the same pitch can be represented by the self-transitions (*e.g.*, $z^\mathrm{L}_{n-1} = z^\mathrm{L}_n = \mathrm{C3}$). On the other hand, for the rhythm guitar part, we need to determine the actual onsets (attack times) in each region of chord symbols $\mathbf{Z}^\mathrm{R}$ because the same chords can be repeated multiple times. To obtain a natural rhythmic pattern within each measure, we perform template matching by using a dictionary of rhythmic patterns (16-dimensional binary vectors). The entries of the dictionary are obtained from a collection of musical pieces. We first obtain a tentative rhythmic pattern by detecting peaks in the activation $\mathbf{H}^\mathrm{R}$ and then use the closest pattern in the dictionary based on the cosine distance for estimating the score.

The variables $\mathbf{Z}$, $\mathbf{U}$, and $\mathbf{H}$ are estimated as follows. First, we generate a sufficient amount of samples of $\mathbf{Z}$, $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{H}$ by using Gibbs sampling. Second, we fix the parameters $\mathbf{W}$ and $\mathbf{H}$ with the last sample and the maximum-a-posteriori (MAP) estimates of $\mathbf{Z}$ and $\mathbf{U}$ are obtained by using the Viterbi algorithm. Finally, the MAP estimate of $\mathbf{H}$ is obtained from the posterior distribution in Eq. (17).

3. EVALUATION

3.1. Experimental Setup

Ten songs by *The Beatles* in 4/4 time consisting of rhythm, lead, and bass guitars, vocals, and drums were used for evaluation. To extract accompaniment sounds by suppressing the drum and vocal sounds, we used a harmonic/percussive source separation method [2] and a singing voice separation method [5]. The tatum-level chord label accuracy was measured for the rhythm guitar and the tatum-level recall and precision rates and F-measure were measured for the lead and bass guitars. The chord vocabulary consisted of 24 labels ($\mathbf{w}^{\mathrm{R}}_{\mathrm{C}}, \ldots, \mathbf{w}^{\mathrm{R}}_{\mathrm{Bm}}$) and the *no chord* regions were not considered. Recall, precision, and F-measure were given by $\mathcal{P} = \mathrm{N}_{\mathrm{tp}}/\mathrm{N}_{\mathrm{sys}}$, $\mathcal{R} = \mathrm{N}_{\mathrm{tp}}/\mathrm{N}_{\mathrm{ref}}$, and $\mathcal{F} = 2\mathcal{R}\mathcal{P}/(\mathcal{R}+\mathcal{P})$, where N_{tp} is the number of correctly-estimated pitches, $\mathrm{N}_{\mathrm{sys}}$ the number of detected pitches, and $\mathrm{N}_{\mathrm{ref}}$ the number of ground-truth pitches.

For comparison, we tested a state-of-the-art method for melody and bass line estimation based on convolutional neural networks (CNNs) [15] (called a *deep saliency method*) for transcribing the lead and bass guitar parts. For transcribing the rhythm guitar part, we tested a state-of-the-art chord estimation method based on CNNs, (Madmom [23]). Since the three guitar parts are estimated at the frame level, the estimated results are quantized in the tatum level. To investigate the sensitivity to initialization of the proposed model, we compared random initialization (FHSMM-RND) and initialization using the results of the above methods (FHSMM-DSM).

The log-frequency magnitude spectrogram of a music signal was obtained by the constant-Q transform [24] with 96 bins/octave, a shifting interval of 11 ms, and a frequency range from 32.7 Hz (C1) to 8372.0 Hz (C9). The tatum times were estimated in advance by using a beat tracking method [23]. The hyperparameters $\mathbf{a}^{\mathrm{R}}_w$ and $\mathbf{b}^{\mathrm{R}}_w$ of the prior on $\mathbf{W}^{\mathrm{R}}$ were determined using the spectra of 24 low-position chords played by an acoustic guitar. The hyperparameters $\mathbf{a}^{\star}_w$ and $\mathbf{b}^{\star}_w$ of the prior on $\mathbf{W}^{\star}$ were determined in a similar way by using the spectra of 45 pitches from E2 to C6 or 28 pitches from E1 to G3. These reference spectra were made by MIDI synthesizers. The hyperparameters of activations were set to $a_{h^{\mathrm{R}}_t} = a_{h^{\star}_t} = 1$, $b_{h^{\mathrm{R}}_t} = b_{h^{\star}_t} = 1$. The uniform initial probabilities were used: $\boldsymbol{\pi}^{\mathrm{L}} = 1/45$, $\boldsymbol{\pi}^{\mathrm{B}} = 1/28$, and $\boldsymbol{\pi}^{\mathrm{R}} = 1/24$. The language model parameters ϕ, ψ, and a dictionary of rhythmic patterns for the rhythm guitar were learned in advance from 206 pieces of Japanese popular musicby the maximum-likelihood method.

3.2. Experimental Results

Table 2 shows the models' performances on multi-guitar transcription. Note that $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ can never be 100% because the lead and bass guitars were assumed to be monophonic here, but this assumption does not hold true in reality, *e.g.*, the lead guitar sometimes plays chords. FHSMM-DSM consistently outperformed the deep saliency method in the transcription of the lead and bass guitars. As shown in Fig. 3, unnatural musical notes such as short insertion errors estimated by the deep saliency method were significantly reduced by the proposed method. This shows the effectiveness of the language models, which induce musical naturalness of estimated scores. FHSMM-RND, on the other hand, underperformed Madmom and the deep saliency method because the NMF-based acoustic model was sensitive to initialization. We found that the proposed method of joint guitar transcription can find more accurate scores than independent transcription methods if it is appropriately initialized. The performance of chord estimation obtained by FHSMM-DSM were almost same as that obtained by Madmom. This is probably because good local optima were already found by Madmom.

Table 2. Performance of transcription for three guitar parts.

Method	Part	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$	Chord
Deep saliency [15]	Lead	32.4	19.0	22.8	–
	Bass	57.0	61.4	58.8	–
Madmom [23]	Rhythm	–	–	–	74.8
FHSMM-RND	Lead	20.8	12.8	15.0	–
	Bass	34.3	36.6	35.3	–
	Rhythm	–	–	–	48.2
FHSMM-DSM	Lead	**34.7**	**19.8**	**24.1**	–
	Bass	**57.7**	**62.2**	**60.0**	–
	Rhythm	–	–	–	74.8

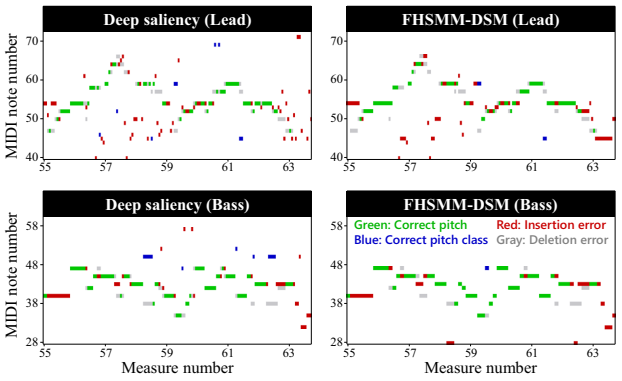

Fig. 3. Musical scores of the lead and bass guitars estimated by the deep saliency method and the proposed method initialized by the deep saliency method for "Don't Bother Me."

Although we achieved an improvement on the transcription accuracy, there is still much room for improving the performance of the lead guitar transcription. First, it is necessary to deal with rests because the lead guitar often has rests during vocal parts. It is also necessary to relax the strong constraints that the lead guitar only plays a monophonic melody and that the rhythm guitar is described with a single set of bases. Another limitation is that the Markov model does not have sufficient capability of expressing the long-term dynamics of musical notes. We plan to extend a deep generative model such as a variational autoencoder (VAE) [25] or a generative adversarial network (GAN) [26] for dealing with time-series data.

4. CONCLUSION

We have described a statistical method to estimate the musical scores of lead, bass, and rhythm guitars from a polyphonic audio signal of typical band-style music. Here, we proposed a unified Bayesian framework for integrating language and acoustic models of multi-instrument music based on a factorial hidden semi-Markov model. We have shown that the unified model improves the accuracy and naturalness of the transcribed scores. We focused on band-style popular music in this paper, but our framework can be applied to various kinds of music transcription involving multi-pitch detection and part assignment. In piano transcription, for example, more reasonable estimations could be obtained by executing the multi-pitch detection and separation of right-hand and left-hand streams jointly [27]. In the future, we plan to develop a system that generates a complete score consisting of vocal, guitar, and drum parts by combining the work described in this paper with automatic vocal (melody) and drum transcription methods.

5. REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri, "Automatic music transcription: Challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[2] Derry Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, 2010, pp. 1–4.

[3] Carl Southall, Ryan Stables, and Jason Hockman, "Automatic drum transcription using bi-directional recurrent neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 591–597.

[4] Chih-Wei Wu and Alexander Lerch, "Drum transcription using partially fixed non-negative matrix factorization with template adaptation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 257–263.

[5] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 745–751.

[6] Emilio Molina, Lorenzo J. Tardón, Ana M. Barbancho, and Isabel Barbancho, "SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 2, pp. 252–263, 2015.

[7] Ryo Nishikimi, Eita Nakamura, Masataka Goto, Katsutoshi Itoyama, and Kazuyoshi Yoshii, "Scale- and rhythm-aware musical note estimation for vocal F0 trajectories based on a semi-tatum-synchronous hierarchical hidden semi-Markov model," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 376–382.

[8] Mert Bay, Andreas F. Ehmann, James W. Beauchamp, Paris Smaragdis, and J. Stephen Downie, "Second fiddle is important too: Pitch tracking individual voices in polyphonic music.," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 319–324.

[9] Graham Grindlay and Daniel P.W. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1159–1169, 2011.

[10] Emmanouil Benetos, Roland Badeau, Tillman Weyde, and Gaël Richard, "Template adaptation for improving automatic music transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 175–180.

[11] Emmanuel Vincent and Xavier Rodet, "Music transcription with ISA and HMM," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2004, pp. 1197–1204.

[12] Emmanuel Vincent, Nancy Bertin, and Roland Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[13] Paris Smaragdis and Judith C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[14] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 5, pp. 927–939, 2016.

[15] Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P. Bello, "Deep salience representations for F0 estimation in polyphonic music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 63–70.

[16] Rodrigo Schramm, Andrew McLeod, Mark Steedman, and Emmanouil Benetos, "Multi-pitch detection and voice assignment for a cappella recordings of multiple singers," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 552–559.

[17] Zoubin Ghahramani and Michael I. Jordan, "Factorial hidden Markov models," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 1996, pp. 472–478.

[18] Gautham J. Mysore and Maneesh Sahani, "Variational inference in non-negative factorial hidden Markov models for efficient audio source separation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012, pp. 1887–1894.

[19] Alexey Ozerov, Cédric Févotte, and Maurice Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 121–124.

[20] Christopher Raphael, "A hybrid graphical model for rhythmic parsing," *Artificial Intelligence*, vol. 137, pp. 217–238, 2002.

[21] Masatoshi Hamanaka, Masataka Goto, Hideki Asoh, and Nobuyuki Otsu, "A learning-based quantization: Unsupervised estimation of the model parameters," in *Proceedings of the International Computer Music Conference (ICMC)*, 2003, pp. 369–372.

[22] Ali Taylan Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Journal of Computational Intelligence and Neuroscience*, , no. 785152, pp. 1–17, 2009.

[23] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer, "Madmom: A new python audio and music signal processing library," in *Proceedings of the ACM International Conference on Multimedia*, 2016, pp. 1174–1178.

[24] Christian Schörkhuber and Anssi Klapuri, "Constant-Q transform toolbox for music processing," in *Proceedings of the Sound and Music Computing Conference (SMC)*, 2010, pp. 3–64.

[25] Diederik P. Kingma and Max Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[27] Eita Nakamura, Kazuyoshi Yoshii, and Shigeki Sagayama, "Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 794–806, 2017.