BAYESIAN DRUM TRANSCRIPTION BASED ON NONNEGATIVE MATRIX FACTOR DECOMPOSITION WITH A DEEP SCORE PRIOR

Shun Ueda   Kentaro Shibata   Yusuke Wada   Ryo Nishikimi   Eita Nakamura   Kazuyoshi Yoshii

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

This paper describes a statistical method of automatic drum transcription that estimates a musical score of bass and snare drums and hi-hats from a drum signal separated from a popular music signal. One of the most effective approaches for this problem is to apply nonnegative matrix factor deconvolution (NMFD) for estimating the temporal activations of drums and then perform thresholding for estimating a drum score. Such a pure audio-based approach, however, cannot avoid musically unnatural scores. To solve this, we propose a unified Bayesian model that integrates an NMFD-based acoustic model evaluating the likelihood of a drum score for a drum spectrogram, with a deep language model serving as a prior (constraint) of the score. The language model can be trained with existing drum scores in the framework of autoencoding variational Bayes and has more expressive power than the conventional statistical models. We derive an inference algorithm using Gibbs sampling, which is a marriage of the solid formalism of Bayesian learning with the expressive power of deep learning. It is shown that the proposed method not only slightly improved the F-measure score but also increased musical naturalness of the transcribed drum scores than NMFD.

***Index Terms*—** Drum transcription, musical language model, NMF, VAE, deep Bayesian learning

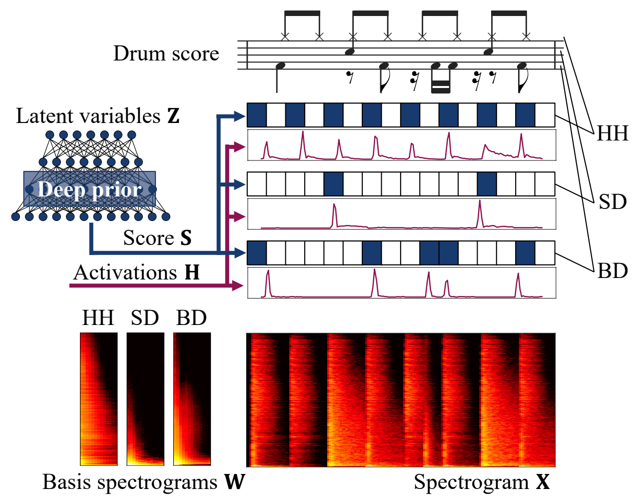

Fig. 1. A hierarchical generative model of a drum-part spectrogram integrating a pretrained DNN-based drum score model (score prior) with an NMFD-based acoustic model (score likelihood). Given a drum-part spectrogram as observed data, a drum score and all variables are estimated by using both models in a Bayesian manner.

1. INTRODUCTION

Automatic drum transcription (ADT) has actively been investigated for describing the rhythmic characteristics of popular music in the field of music information retrieval (MIR) [1]. Although many different types of drum instruments such as floor, low, and high toms and ride and crash cymbals are included in a drum kit, three kinds of drum instruments, *i.e.*, bass and snare drums and hi-hats, have commonly been focused on because they form the rhythmic backbone of popular music. Most studies on ADT aim to estimate *drum rolls* describing the onset times of those drums in a similar way that most studies on automatic music transcription (AMT) aim to estimate *piano rolls* describing the onset and offset times of pitched musical instruments. To complete ADT, it is thus necessary to convert drum rolls to drum scores by quantizing the onset times of the drums. Such a process is called rhythm transcription in AMT [2, 3], but this has scarcely been investigated in ADT.

A typical approach to ADT is to use nonnegative matrix factorization (NMF) for decomposing a drum spectrogram into the basis spectra and temporal activations of the three drums [4–6]. NMF has often been used for AMT and is especially suitable for ADT because drum sounds appear repeatedly with different combinations and volumes and the magnitude spectrogram of a drum part can thus be approximated as a low-rank matrix. Since the acoustic characteristics of each drum cannot be fully represented by a basis *spectrum*, Smaragdis [7] proposed a convolutional extension of NMF called nonnegative matrix factor decomposition (NMFD) that approximates a drum-part spectrogram as a patchwork consisting of overlapping basis *spectrograms* of the drums. To detect the onset times of the drums, simple peak-picking or thresholding is typically applied to the estimated activations. To avoid such a separate post-processing, Liang *et al.* [8] proposed beta-process NMF (BP-NMF) that introduces binary variables (masks) describing the presence or absence of basis components at each time.

Although NMF and its variants have been used successfully for ADT, musically unnatural drum rolls are often obtained. If a dictionary of drum patterns is available, one can categorize each segment of the estimated drum rolls into one of the registered patterns [9]. This approach, however, cannot deal with unregistered drum patterns. Recurrent neural networks (RNNs) have recently been used for learning direct conversion of a drum-part spectrogram to a drum roll in a supervised manner and significantly improved the performance [10, 11]. However, musically unnatural scores cannot be avoided because RNNs are used for learning the temporal dynamics of drum sound mixtures at the *frame level* and those of drum scores at the *tatum level* are not considered.

The limitation of these pure acoustic models calls for a music language model defined on symbolic musical scores. Such language models have recently been used successfully for AMT [12–14]. A

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

basic approach to representing the sequential dependency of musical notes is to use first- or lower-order Markov models or hidden Markov models (HMMs) [13]. The expressive power of these models, however, is severely limited and higher-order models are computationally prohibitive. RNN-based language models have recently been proposed to learn long-term dependency of musical notes and used for estimating musical scores from piano rolls estimated by an NMF-like low-rank acoustic model [14]. Principled integration of a language model defined on discrete symbols and an acoustic model defined on continuous values is still an open problem.

In this paper, we propose a new approach to ADT based on a unified Bayesian model integrating a DNN-based language model with an NMFD-based acoustic model (Fig. 1) under an assumption that tatum times (16th-note-level beat times) and bar lines are given in advance (*e.g.* by a beat tracking method [15]). The acoustic model evaluates the likelihood of a drum score (tatum-level binary variables) for a drum spectrogram and the language model evaluates the prior probability (musical appropriateness) of the score. While the physical additivity of drum sounds can be represented well by a linear model based on NMFD, the complicated syntactic structures of drum scores are hard to be explicitly represented. We thus use a variational autoencoder (VAE) [16] for learning an implicit generative model of one-measure drum patterns with their latent feature representations from existing drum patterns in an unsupervised manner. Given a drum spectrogram, a drum score (a sequence of one-measure drum patterns) and all variables of the language and acoustic models can be estimated in a principled manner via Gibbs sampling.

At the heart of this study is a marriage of the solid formalism of Bayesian learning with the expressive power of deep learning. This is the first attempt that utilizes a powerful deep prior model for ADT and that can be applied to more general types of AMT. A key advantage of our deep Bayesian approach is that a huge amount of drum patterns available on the Web [17] can be used for learning the language model, while a standard approach to end-to-end learning needs time-aligned pair data for supervised learning.

2. PROPOSED METHOD

This section describes the proposed method that estimates a drum score from a drum-part signal separated from a popular music signal using harmonic/percussive source separation (HPSS) [18].

2.1. Problem Specification

The problem of ADT is formalized as follows:

Input: The magnitude spectrogram of a target signal $\mathbf{X} \in \mathbb{R}_+^{F \times T}$ with 16th-note-level tatum times and bar lines
Output: Drum score $\mathbf{S} \in \{0,1\}^{K \times R}$.

Here, F is the number of frequency bins, T the number of time frames, $K = 3$ the number of drum instruments (snare and bass drums and hi-hats), and R the number of tatums in the observed signal. The target signal is assumed to include only the percussive components obtained from the HPSS method [18]. The binary mask S_{kr} indicates whether drum k has an onset at tatum r. Note that $\mathbf{S}$ can be divided into measures (drum patterns).

2.2. Model Formulation

We formulate a hierarchical generative model of a magnitude spectrogram $\mathbf{X}$ by integrating a DNN-based language model of binary masks $\mathbf{S}$ with an NMFD-based acoustic model of $\mathbf{X}$ (Fig. 1).

2.2.1. NMFD-Based Acoustic Model (Score Likelihood)

The magnitude spectrogram $\mathbf{X}$ is approximated by using basis spectrograms $\mathbf{W} \in \mathbb{R}_+^{(K+1) \times F \times M}$, activation vectors $\mathbf{H} \in \mathbb{R}_+^{(K+1) \times T}$, and binary masks $\mathbf{S} \in \{0,1\}^{K \times R}$ as follows:

$$X_{ft} \approx Y_{ft} \stackrel{\text{def}}{=} \sum_{m=1}^{M} \sum_{k=0}^{K} Y_{ftkm}. \tag{1}$$

Here, Y_{ftkm} is given by

$$\begin{cases} Y_{ftkm} = W_{kfm} H_{k,t-m} S_{k,r(t-m)} & (k \geq 1), \\ Y_{ft0m} = W_{0fm} H_{0,t-m}, \end{cases} \tag{2}$$

where M is the number of frames forming each basis spectrogram, $\{W_{kfm}\}_{f=1}^{F}$ $(k \geq 1)$ is the basis spectrum of drum k at frame m and $r(t)$ denotes the tatum to which frame t belongs. We have introduced an additional basis spectrogram W_{0fm} and an activation vector H_{0t} to represent possible noise added to the target drum sounds. To evaluate the approximation error of Eq. (1), we use the Kullback-Leibler (KL) divergence as in KL-NMF [19]. In terms of probabilistic modeling, the minimization of the KL divergence is equivalent to the maximization of the Poisson likelihood given by

$$X_{ft} \sim \mathrm{Poisson}(Y_{ft}). \tag{3}$$

To complete Bayesian formulation, we put conjugate gamma priors on $\mathbf{W}$ as follows:

$$\begin{cases} W_{kfm} \sim \mathrm{Gamma}(a_{kfm}, b_{kfm}) & (k \geq 1), \\ W_{0fm} \sim \mathrm{Gamma}(a_0, b_0), \end{cases} \tag{4}$$

where $\mathrm{Gamma}(a_*, b_*)$ denotes a gamma distribution with shape and rate hyperparameters a_* and b_*. Similarly, we put conjugate gamma priors $\mathbf{H}$ as follows.

$$\begin{cases} H_{kt} \sim \mathrm{Gamma}(c_k, d_k) & (k \geq 1), \\ H_{0t} \sim \mathrm{Gamma}(c_0, d_0), \end{cases} \tag{5}$$

where c_k, d_k c_0, and d_0 are hyperparameters.

2.2.2. DNN-Based Language Model (Score Prior)

The binary masks $\mathbf{S}$ are assumed to independently follow Bernoulli distributions as follows:

$$S_{kr} \sim \mathrm{Bernoulli}(\pi_{kr}), \tag{6}$$

where π_{kr} indicates the prior probability of the presence of the onset of drum k at tatum r. For mathematical convenience, we rewrite the drum- and tatum-wise representation given by Eq. (6) as a measure-wise representation as follows:

$$\mathbf{s}_i \sim \mathrm{Bernoulli}(\boldsymbol{\pi}_i), \tag{7}$$

where $\mathbf{s}_i$ and $\boldsymbol{\pi}_i$ are $16K$-dimensional binary and real-valued vectors consisting of S_{kr}'s and π_{kr}'s in measure i $(0 \leq i \leq I-1)$, respectively. The core part of the proposed method is that $\boldsymbol{\pi}_i$ is represented by an implicit deep generative model as follows:

$$\mathbf{z}_i \sim \mathcal{N}(0,1), \tag{8}$$

$$\boldsymbol{\pi}_i = \mathrm{DNN}_\theta(\mathbf{z}_i), \tag{9}$$

where DNN_θ is a non-linear function with parameters θ that maps $\mathbf{z}_i$ to $\boldsymbol{\pi}_i$ and $\mathbf{z}_i$ is a V-dimensional latent representation of the drum pattern of measure i. The deep score prior $p_\theta(\mathbf{S})$ is obtained by marginalizing out the latent variables $\mathbf{Z}$ from the implicit generative model given by $p_\theta(\mathbf{S}|\mathbf{Z})p(\mathbf{Z})$.

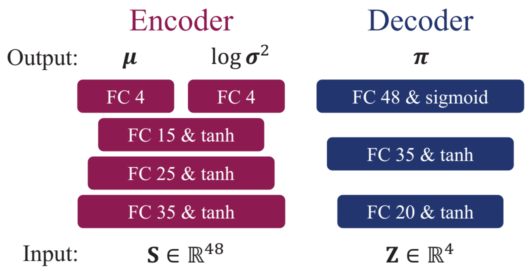

Fig. 2. The VAE of one-measure drum patterns.

2.3. Score Prior Learning

To estimate the deep score prior $p_\theta(\mathbf{S})$, we train a variational autoencoder (VAE) for existing drum patterns $\mathbf{S}$ in an unsupervised manner. Our goal is to estimate the DNN parameters θ that maximize the likelihood given by $p_\theta(\mathbf{S})$. Since the direct maximization of $p_\theta(\mathbf{S})$ is intractable, we derive the lower bound of $\log p_\theta(\mathbf{S})$ that can be maximized easily. More specifically, introducing an arbitrary variational distribution $q(\mathbf{Z})$ and using Jensen's inequality, the lower bound of $\log p_\theta(\mathbf{S})$ can be derived as follows:

$$\log p_\theta(\mathbf{S}) \geq -\mathbb{KL}[q(\mathbf{Z})|p(\mathbf{Z})] + \mathbb{E}_q[\log p_\theta(\mathbf{S}|\mathbf{Z})]. \tag{10}$$

As an instance of $q(\mathbf{Z})$, we formulate a *recognition* model $q_\phi(\mathbf{Z}|\mathbf{S})$ with parameters ϕ defined as follows:

$$q_\phi(\mathbf{Z}|\mathbf{S}) = \prod_{i=0}^{I-1} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_\phi(\mathbf{s}_i), \boldsymbol{\sigma}^2_\phi(\mathbf{s}_i)), \tag{11}$$

where $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\sigma}^2_\phi$ are nonlinear functions defined with DNNs whose input and output are $16K$- and V-dimensional vectors, respectively. The lower bound of $\log p_\theta(\mathbf{S})$ can be further written as follows:

$$\begin{aligned}
&\log p_\theta(\mathbf{S}) \\
&\geq \frac{1}{2}\sum_{i,v}\left(1 + \log \boldsymbol{\sigma}^2_{\phi,v}(\mathbf{s}_i) - \boldsymbol{\mu}^2_{\phi,v}(\mathbf{s}_i) - \boldsymbol{\sigma}^2_{\phi,v}(\mathbf{s}_i)\right) \\
&\quad + \sum_{k,r}\mathbb{E}_q[S_{kr}\log \pi_{kr} + (1-S_{kr})\log(1-\pi_{kr})],
\end{aligned} \tag{12}$$

where $\boldsymbol{\mu}^2_{\phi,v}(\mathbf{s}_i)$ is the vth dimension of the V-dimensional output of $\boldsymbol{\mu}^2_\phi(\mathbf{s}_i)$ and $\boldsymbol{\sigma}^2_{\phi,v}(\mathbf{s}_i)$is defined similarly. Eq. (12) is a function of θ and ϕ because $\boldsymbol{\pi}$ is determined by Eq. (9). Both θ and ϕ are jointly optimized such that the lower bound given by Eq. (12) is maximized by a stochastic gradient descent method such as Adam [20].

2.4. Score Posterior Computation

Given $\mathbf{X}$ as observed data, we aim to compute the posterior distribution $p(\mathbf{W}, \mathbf{H}, \mathbf{S}, \mathbf{Z}|\mathbf{X})$. Since this cannot be calculated analytically, we use Gibbs sampling for iteratively and alternately updating $\mathbf{W}$, $\mathbf{H}$, $\mathbf{S}$, and $\mathbf{Z}$ in a stochastic manner.

2.4.1. Updating Drum Score

Using the acoustic model with $\mathbf{W}$ and $\mathbf{H}$ and the language model with $\mathbf{Z}$, binary masks $\mathbf{S}$ are sampled as follows:

$$S_{kr} \sim \text{Bernoulli}\left(\frac{\mathrm{P}^1_{kr}}{\mathrm{P}^0_{kr} + \mathrm{P}^1_{kr}}\right), \tag{13}$$

$$\mathrm{P}^0_{kr} \propto (1-\pi_{kr})\, p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 0), \tag{14}$$

$$\mathrm{P}^1_{kr} \propto \pi_{kr}\, p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 1), \tag{15}$$

where the first and second terms of Eq. (14) or Eq. (15) indicate the prior probability and the acoustic likelihood, respectively, and $\mathbf{S}_{\neg(kr)}$ denotes the subset of $\mathbf{S}$ excluding S_{kr}. Note that $\boldsymbol{\pi}$ depends on $\mathbf{Z}$. The likelihood terms of Eq. (14) and Eq. (15) are given by

$$\begin{aligned}
&p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 0) \\
&= \prod_{t\in\{r(t)=r\}}\prod_f \left(Y^{\neg k}_{ft} + \sum_m W_{kfm}H_{k,t-m}\right)^{X_{ft}} \\
&\qquad\qquad \cdot \exp\left\{-\sum_m W_{kfm}H_{k,t-m}\right\},
\end{aligned} \tag{16}$$

$$\begin{aligned}
&p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{S}_{\neg(kr)}, S_{kr} = 1) \\
&= \prod_{t\in\{r(t)=r\}}\prod_f \left(Y^{\neg k}_{ft}\right)^{X_{ft}},
\end{aligned} \tag{17}$$

where $Y^{\neg k}_{ft}$ is given by

$$Y^{\neg k}_{ft} = \sum_{l\neq k}\sum_m Y_{ftkm} \quad (k \geq 1). \tag{18}$$

2.4.2. Updating NMFD-Based Acoustic Model

To sample $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{S}$ involved in Bayesian NMFD with binary masks $\mathbf{S}$, we extend a Gibbs sampling method proposed for Bayesian NMF with binary masks called BP-NMF [8]. More specifically, conditioned by $\mathbf{H}$ and $\mathbf{S}$, $\mathbf{W}$ is sampled as follows:

$$W_{kfm} \sim \text{Gamma}(\hat{a}_{kfm}, \hat{b}_{kfm}), \tag{19}$$

$$\begin{cases} \hat{a}_{kfm} = \sum_t X_{ft}\lambda_{ftkm} + a_{kfm} & (k \geq 1), \\ \hat{a}_{0fm} = \sum_t X_{ft}\lambda_{ft0m} + a_0, \end{cases} \tag{20}$$

$$\begin{cases} \hat{b}_{kfm} = \sum_t H_{k,t-m}S_{k,t-m} + b_{kfm} & (k \geq 1), \\ \hat{b}_{0fm} = \sum_t H_{0,t-m} + b_0, \end{cases} \tag{21}$$

where λ_{ftkm} is an auxiliary variable given by

$$\lambda_{ftkm} = \frac{Y_{ftkm}}{Y_{ft}}. \tag{22}$$

Similarly, conditioned by $\mathbf{W}$ and $\mathbf{S}$, $\mathbf{H}$ is sampled as follows:

$$H_{kt} \sim \text{Gamma}(\hat{c}_{kt}, \hat{d}_{kt}), \tag{23}$$

$$\begin{cases} \hat{c}_{kt} = \sum_{f,m} X_{ft}\lambda_{f,t+m,km} + c_k & (k \geq 1), \\ \hat{c}_{0t} = \sum_{f,m} X_{ft}\lambda_{f,t+m,0m} + c_0, \end{cases} \tag{24}$$

$$\begin{cases} \hat{d}_{kt} = \sum_{f,m} W_{kfm}S_{kt} + d_k & (k \geq 1), \\ \hat{d}_{0t} = \sum_{f,m} W_{0fm} + d_0. \end{cases} \tag{25}$$

2.4.3. Updating DNN-Based Language Model

Since it is difficult to analytically calculate the posterior distribution of $\mathbf{Z}$, we use a Metropolis-Hastings method to update $\mathbf{Z}$. A proposal of $\mathbf{z}^*_i$ at each bar i is sampled in a way of random walk by using a Gaussian distribution as follows:

$$\mathbf{z}^*_i \sim q(\mathbf{z}^*_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i, 0.1). \tag{26}$$

The proposal $\mathbf{z}^*_i$ is accepted as the next $\mathbf{z}_i$ with the following acceptance rate $a_{\mathbf{z}^*_i|\mathbf{z}_i}$:

$$a_{\mathbf{z}^*_i|\mathbf{z}_i} = \min\left(1, \frac{p(\mathbf{z}^*_i)}{p(\mathbf{z}_i)}\prod_{k,r\in\{\text{bar}(r)=i\}}\frac{p(S_{kr}|\mathbf{z}^*_i)}{p(S_{kr}|\mathbf{z}_i)}\right). \tag{27}$$

Method	Part	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$
NMFD	HH	79.4	60.9	69.0
	SD	63.2	63.6	63.4
	BD	82.3	80.2	81.2
VAE-NMFD	HH	80.9	61.4	69.8
	SD	67.6	65.4	66.5
	BD	83.0	79.4	81.2

Table 1. Performances of ADT for RWC popular music database. The "HH", "SD", and "BD" represent the hi-hats and snare and bass drums, respectively.

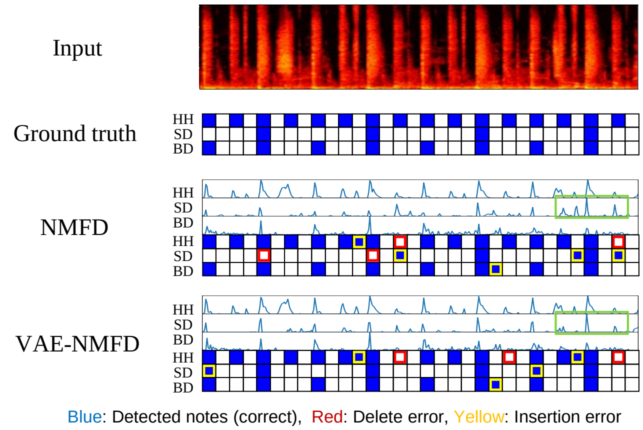

Fig. 3. Examples of drum scores estimated by NMFD (baseline) and VAE-NMFD (proposed). For VAE-NMFD, activations obtained after applying the masks are shown.

Here, $\mathrm{bar}(r) = \lfloor \frac{r}{16} \rfloor$ denotes the measure to which tatum r belongs. To estimate $\mathbf{Z}$ effectively, we initialize $\mathbf{Z}$ with samples drawn from the recognition model $q_\phi(\mathbf{Z}|\mathbf{S})$ with initial estimates of $\mathbf{S}$.

3. EVALUATION

3.1. Experimental Setup

For evaluation, we used audio signals in the RWC popular music database [21]. Those signals were converted into monaural signals and divided into segments of 30-second length. The second segment of each piece was used for evaluation. We selected 64 pieces in which bass and snare drums and hi-hats are played at least once. We split the selected audio signals into segments of 1 measure using tatum times obtained from the annotations [22]. The tatum times we used were shifted 0.03 seconds earlier from the original annotations to align them with the onset times of the drum sounds.

All songs were sampled at 44.1 kHz, and we obtained magnitude spectrograms using an STFT with a Hann window of 2048 points and a shifting interval of 441 points (10 ms). Moreover, we applied HPSS [18] for the spectrograms to separate the drum part spectrograms. Each magnitude spectrogram was normalized so that the average magnitude becomes unity.

To determine the hyperparameters a_{kfm} and b_{kfm} $(k \geq 1)$, we used the isolated sounds of bass and snare drums and hi-hats from the RWC musical instrument sound database [23]. The magnitude spectrograms of those sounds were obtained similarly as above and the template spectrogram of each drum was estimated by applying NMFD with a single basis to the spectrogram obtained by concatenating all the spectrograms of the target drum. The hyperparameters were set so that the means of the Gamma priors were the same as the basis spectrograms and the variance of the Gamma priors was 0.01. The other hyperparameters of the priors on basis spectrograms $\mathbf{W}$ and activation vectors $\mathbf{H}$ were set as $a_{0fm} = 0.05$, $b_{0fm} = 50.0$, $c_0 = 50.0$, $d_0 = 50.0$, $c_k = 1.0$, and $d_k = 50.0$. We trained the VAE network using 41474 bars obtained from drum scores of The Beatles and Japanese popular music, which had no overlaps with the test data. The number of frames forming each basis spectrogram was $M = 20$. The dimension of the latent variable $\mathbf{z}_i$ was $V = 4$.

Performance of ADT was measured by the precision and recall rates and F-measure defined as follows:

$$\mathcal{P} = \frac{N_c}{N_e}, \quad \mathcal{R} = \frac{N_c}{N_g}, \quad \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}}, \tag{28}$$

where N_e, N_g, and N_c are the numbers of estimated, ground-truth, and correct notes, respectively. For each k (≥ 1), note onsets t^* are detected using the estimated from $\mathbf{H}$ and $\mathbf{S}$ by the following conditions for forming a peak:

$$H_{kt^*} S_{kt^*} \geq 0.3 \cdot \max_t \{H_{kt} S_{kt}\}, \tag{29}$$

$$H_{kt^*} S_{kt^*} = \max_{t^*-5 \leq t \leq t^*+5} \{H_{kt} S_{kt}\}. \tag{30}$$

When the time difference between an estimated note and a ground-truth note was within 50 ms, we judged the estimated note as correct.

3.2. Experimental Results

The experimental results of ADT are shown in Table 1. For snare drum and hi-hats, the proposed method significantly outperformed NMFD in all the metrics. For bass drum, the recall rate for the proposed method was slightly worse than that of NMFD and the F-measure was even. In the example in Fig. 3, the snare drum part obtained by NMFD (acoustic model) had unnatural rhythmic patterns (for example in the last half measure) whereas that obtained by the proposed method was musically natural. These results indicate that the proposed method integrating the DNN-based language model and the NMFD-based acoustic model not only improved the objective evaluation metrics but also increased the musical naturalness of the transcribed scores. These results clearly demonstrate the effectiveness of the proposed method.

4. CONCLUSION

This paper has presented a statistical method of ADT that integrates an NMFD-based acoustic model with a VAE-based deep language model in a unified Bayesian manner. A key advantage of our deep Bayesian approach is that the language model can be learned from musical scores, while a standard approach to end-to-end learning needs time-aligned pair data for supervised learning. This approach can be applied to more general types of music transcription. The experimental results showed that the proposed method can estimate musically natural scores by leveraging the powerful deep score prior.

A future direction is to integrate the present method with a statistical method of beat and downbeat detection for joint estimation of drum scores and beat times, similarly as in [24]. We also plan to represent the temporal dependency and repetitive structures of drum patterns by using a time-series or recurrent extension of the VAE.

5. REFERENCES

[1] C. Wu, C. Dittmar, C. Southall, and R. Vogl, "A review of automatic drum transcription," *Journal of IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1457–1483, 2018.

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

[3] E. Nakamura, K. Yoshii, and S. 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.

[4] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *13th European Signal Processing Conference, (EUSIPCO)*, 2005, pp. 1–4.

[5] C. Dittmar and D. Gärtner, "Real-time transcription and separation of drum recordings based on NMF decomposition," in *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, 2014, pp. 187–194.

[6] C. Wu and A. 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.

[7] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2004, pp. 494–499.

[8] D. Liang, M. D. Hoffman, and D. P. W. Ellis, "Beta process sparse nonnegative matrix factorization for music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 375–380.

[9] L. Thompson, S. Dixon, and M. Mauch, "Drum transcription via classification of bar-level rhythmic patterns," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 187–192.

[10] C. Southall, R. Stables, and J. 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.

[11] R. Vogl, M. Dorfer, and P. Knees, "Drum transcription from polyphonic music with recurrent neural networks," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 201–205.

[12] S. Sigtia, E. Benetos, and S. 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.

[13] R. Nishikimi, E. Nakamura, M. Goto, K. Itoyama, and K. 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.

[14] A. Ycart and E. Benetos, "Polyphonic music sequence transduction with meter-constrained LSTM networks," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[15] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "Madmom: A new python audio and music signal processing library," in *Proceedings of the ACM International Conference on Multimedia*, 2016, pp. 1174–1178.

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

[17] C. Raffel, "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching", Ph.D. thesis, COLUMBIA UNIVERSITY, 2016.

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

[19] P. Smaragdis and J. 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.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, vol. 1412.6980, 2014.

[21] M. Goto, H. Hashiguchi, H. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

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

[23] M. Goto, H. Hashiguchi, H. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2003, pp. 229–230.

[24] R. Vogl, M. Dorfer, G. Widmer, and P. Knees, "Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 150–157.