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

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 spectrograms 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].
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 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

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 \( X \in \mathbb{R}_+^{F \times T} \) with 16th-note-level tatum times and bar lines

Output: Drum score \( S \in \{0, 1\}^{K \times R} \).

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 tums 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 \( S \) can be divided into measures (drum patterns).

2.2. Model Formulation

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

2.2.1. NMFD-Based Acoustic Model (Score Likelihood)

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

\[
X_{ft} \approx Y_{ft} \triangleq \sum_{m=1}^M \sum_{k=0}^K Y_{ftkm}.
\]

Here, \( Y_{ftkm} \) is given by

\[
Y_{ftkm} = W_{kfm} H_{k,t-m} S_{kr,(t-m)} \quad (k \geq 1),
\]

\[
Y_{ft0m} = W_{0fm} H_{0,t-m}.
\]

where \( M \) is the number of frames forming each basis spectrogram, \( \{W_{kfm}\}_{k=1}^M \) 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 \( H_{0m} \) 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

\[
X_{ft} \sim \text{Poisson}(Y_{ft}).
\]

To complete Bayesian formulation, we put conjugate gamma priors on \( W \) as follows:

\[
W_{kfm} \sim \text{Gamma}(a_{kfm}, b_{kfm}) \quad (k \geq 1),
\]

\[
W_{0fm} \sim \text{Gamma}(a_0, b_0),
\]

where \( \text{Gamma}(a, b) \) denotes a gamma distribution with shape and rate hyperparameters \( a \) and \( b \). Similarly, we put conjugate gamma priors \( H \) as follows:

\[
H_{kt} \sim \text{Gamma}(c_k, d_k) \quad (k \geq 1),
\]

\[
H_{0t} \sim \text{Gamma}(c_0, d_0),
\]

where \( c_k, d_k, c_0, \) and \( d_0 \) are hyperparameters.

2.2.2. DNN-Based Language Model (Score Prior)

The binary masks \( S \) are assumed to independently follow Bernoulli distributions as follows:

\[
S_{kr} \sim \text{Bernoulli}(\pi_{kr}),
\]

where \( \pi_{kr} \) 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:

\[
s_i \sim \text{Bernoulli}(\pi_i),
\]

where \( s_i \) and \( \pi_i \) are 16\( K \)-dimensional binary and real-valued vectors consisting of \( S_{kr} \)'s and \( \pi_{kr} \)'s in measure \( i \) (0 ≤ \( i \) ≤ 1 − 1), respectively. The core part of the proposed method is that \( \pi_i \) is represented by an implicit deep generative model as follows:

\[
z_i \sim N(0, 1),
\]

\[
\pi_i = \text{DNN}_\theta(z_i),
\]

where \( \text{DNN}_\theta \) is a non-linear function with parameters \( \theta \) that maps \( z_i \) to \( \pi_i \) and \( z_i \) is a \( V \)-dimensional latent representation of the drum pattern of measure \( i \). The deep score prior \( p_\theta(S) \) is obtained by marginalizing out the latent variables \( Z \) from the implicit generative model given by \( p_\theta(S|Z)p(Z) \).
To estimate the deep score prior \( p_{\theta}(S) \), we train a variational autoencoder (VAE) for existing drum patterns \( S \) in an unsupervised manner. Our goal is to estimate the DNN parameters \( \theta \) that maximize the likelihood given by \( p_{\theta}(S) \). Since the direct maximization of \( p_{\theta}(S) \) is intractable, we derive the lower bound of \( \log p_{\theta}(S) \) that can be maximized easily. More specifically, introducing an arbitrary variational distribution \( q(Z) \) and using Jensen’s inequality, the lower bound of \( \log p_{\theta}(S) \) can be derived as follows:

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

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

\[
q_{\theta}(Z|S) = \prod_{i=0}^{l-1} \mathcal{N}(z_i|\mu_{\phi}(s_i), \sigma^2_{\phi}(s_i)),
\]

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

\[
\log p_{\theta}(S) \geq \frac{1}{2} \sum_{k,v} \left( 1 + \log \sigma^2_{\phi,v}(s_i) - \mu^2_{\phi,v}(s_i) - \sigma^2_{\phi,v}(s_i) \right) + \mathbb{E}_q[S_{kr}] \log \pi_{kr} + (1 - S_{kr}) \log(1 - \pi_{kr}),
\]

where \( \mu_{\phi,v}(s_i) \) is the \( v \)-th dimension of the \( V \)-dimensional output of \( \mu_{\phi}(s_i) \) and \( \sigma^2_{\phi,v}(s_i) \) is defined similarly. Eq. (12) is a function of \( \theta \) and \( \phi \) because \( \pi \) is determined by Eq. (9). Both \( \theta \) and \( \phi \) 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 \( X \) as observed data, we aim to compute the posterior distribution \( p(W, H, S, Z|X) \). Since this cannot be calculated analytically, we use Gibbs sampling for iteratively and alternately updating \( W, H, S, \) and \( Z \) in a stochastic manner.

#### 2.4.1. Updating Drum Score

Using the acoustic model with \( W \) and \( H \) and the language model with \( Z \), binary masks \( S \) are sampled as follows:

\[
S_{kr} \sim \text{Bernoulli} \left( \frac{P^1_{kr}}{P^1_{kr} + P^0_{kr}} \right),
\]

\[
P^0_{kr} \propto (1 - \pi_{kr}) \cdot p(X|W, H, S_{(kr)}, S_{kr} = 0),
\]

\[
P^1_{kr} \propto \pi_{kr} \cdot p(X|W, H, S_{(kr)}, S_{kr} = 1),
\]

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

\[
p(X|W, H, S_{(kr)}, S_{kr} = 0) = \prod_{t \in \{r(t)=r\}} \left( \sum_m W_{km} \cdot H_{kr,t-m} \right)^{X_{ft}},
\]

\[
p(X|W, H, S_{(kr)}, S_{kr} = 1) = \prod_{t \in \{r(t)=r\}} Y_{ft}^{X_{ft}},
\]

where \( Y_{ft}^{X_{ft}} \) is given by

\[
Y_{ft}^{X_{ft}} = \sum_{l \neq k} Y_{ftkm} (k \geq 1).
\]

#### 2.4.2. Updating NMFD-Based Acoustic Model

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

\[
W_{km} \sim \text{Gamma}(\hat{a}_{km}, \beta_{km}),
\]

\[
\{ \hat{a}_{km} = \sum_t X_{ft} \lambda_{ftkm} + a_{km} \} \quad (k \geq 1),
\]

\[
\{ \beta_{km} = \sum_t H_{km} S_{tl,t-m} + b_{km} \} \quad (k \geq 1),
\]

where \( \lambda_{ftkm} \) is an auxiliary variable given by

\[
\lambda_{ftkm} = \frac{Y_{ftkm}}{Y_{ft}}.
\]

Similarly, conditioned by \( W \) and \( S, H \) is sampled as follows:

\[
H_{km} \sim \text{Gamma}(\hat{c}_k, \hat{d}_k),
\]

\[
\{ \hat{c}_k = \sum_{f,m} X_{ft} \lambda_{f,t+m,k,m} + c_k \} \quad (k \geq 1),
\]

\[
\{ \hat{d}_k = \sum_{f,m} W_{km} S_{tk} + d_k \} \quad (k \geq 1),
\]

\[
\{ \hat{d}_f = \sum_{m} W_{km} S_{tk} + d_0 \}.
\]

#### 2.4.3. Updating DNN-Based Language Model

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

\[
z_i^* \sim q(z_i^*|z_i) = \mathcal{N}(z_i, 0.1).
\]

The proposal \( z_i^* \) is accepted as the next \( z_i \) with the following acceptance rate \( a_{z_i^*|z_i} \):

\[
a_{z_i^*|z_i} = \min \left( 1, \frac{p(z_i^*)}{p(z_i)} \prod_{k,r \in \{\text{bar}(r)=i\}} \frac{p(S_{kr}|z_i^*)}{p(S_{kr}|z_i)} \right).
\]
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 patterns by using a time-series or recurrent extension of the VAE. We also planned to represent the temporal dependency and repetitive structures of drum scores and beat times, similarly as in [24]. We also plan to estimate Z effectively, we initialize Z with samples drawn from the recognition model qφ(Z|S) with initial estimates of S.

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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Part</th>
<th>P(%)</th>
<th>R(%)</th>
<th>F(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMFD</td>
<td>HH</td>
<td>79.4</td>
<td>60.9</td>
<td>69.0</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>63.2</td>
<td>63.6</td>
<td>63.4</td>
</tr>
<tr>
<td></td>
<td>BD</td>
<td>82.3</td>
<td>80.2</td>
<td>81.2</td>
</tr>
<tr>
<td>VAE-NMFD</td>
<td>HH</td>
<td>80.9</td>
<td>61.4</td>
<td>69.8</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>67.6</td>
<td>65.4</td>
<td>66.5</td>
</tr>
<tr>
<td></td>
<td>BD</td>
<td>83.0</td>
<td>79.4</td>
<td>81.2</td>
</tr>
</tbody>
</table>

Ground truth

HH
SD
BD
HH
SD
BD
NMFD
HH
SD
BD
HH
SD
BD
VAE-NMFD
HH
SD
BD
HH
SD
BD

Input

Ground truth

HH
SD
BD
HH
SD
BD

NMFD

HH
SD
BD
HH
SD
BD

VAE-NMFD

HH
SD
BD
HH
SD
BD

Blue: Detected notes (correct), Red: Delete error, Yellow: Insertion error

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.

conditions for forming a peak:

\[ H_{kt+}, S_{kt+} \geq 0.3 \cdot \max_t \{ H_{kt} S_{kt}\}, \]

\[ H_{kt+}, S_{kt+} = \max_{t' - 5 \leq t' \leq t + 5} \{ H_{kt} S_{kt}\}. \]
5. REFERENCES


