2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Joint Drum Transcription and Metrical Analysis Based on Periodicity-Aware Multi-Task Learning

Daichi Kamakura, Eita Nanamura, Takehisa Oyama, and Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Japan
E-mail: {kamakura, nakamura, ooyama}@sap.ist.i.kyoto-u.ac.jp, yoshii@i.kyoto-u.ac.jp

***Abstract*—This paper describes a multi-task learning method that detects the onset times of drums and the beat and downbeat times from a music signal. Since the drum part typically consists of repetitive patterns synchronized with the metrical structure in popular music, drum transcription and metrical analysis would benefit each other. The basic approach is to train a deep neural network (DNN) with a branching architecture that extracts latent features common to both tasks from a music signal and then feeds them to task-specific networks separately. The estimated frame-level posterior probabilities of drum onsets, beats, and downbeats, however, often have weak and inconsistent periodic structures. To solve this problem, we propose a regularized training method that encourages the three probability sequences to be highly auto- and cross-correlated. Specifically, our method aims to minimize the entropy of the auto-spectrum computed from each probability sequence and that of the cross-spectrum computed from each of the three sequence pairs, because a smaller entropy indicates a stronger element-wise periodicity or pairwise consistency. The experiment showed the mutual benefit of drum transcription and metrical analysis and the effectiveness of the periodicity- and consistency-based regularizations.**

I. Introduction

Automatic drum transcription (ADT) is one of the most fundamental subtasks of automatic music transcription (AMT) that aims to estimate the symbolic score of a music signal. Since the drum part affects the rhythmic characteristics of the song in popular music, drum transcription forms the basis of music information retrieval (MIR). Although the ultimate goal of ADT is audio-to-score transcription, we focus on audio-to-MIDI transcription that aims to estimate the onset times of drums in seconds as an intermediate goal.

In recent studies on audio-to-MIDI drum transcription, deep neural networks (DNNs) have successfully been used for estimating from a music spectrogram the posterior probabilities of the presence of drum onsets at the frame level [1]–[3]. Given a ground-truth binary sequence representing the presence or absence of drum onsets, one can train a DNN such that the posterior probability of the binary sequence is maximized, i.e., the cross-entropy loss is minimized. If only a limited amount of training data is available as is often the case with AMT [4], [5], however, the repetitive structures of the drum part observed at multiple metrical levels might not be captured well.

In this paper, we take the multi-task learning approach to joint drum transcription and metrical analysis (beat and downbeat detection). Both tasks are typically defined as binary classification problems that aim to estimate the presence or absence of drum onsets, beats, and downbeats at the frame level, and are expected to benefit each other by extracting common latent features with respect to periodicity from a music signal [2]. In this approach, one may train a DNN with a common feature extractor followed by task-specific binary classifiers. In such a basic branching architecture, however, the posterior probabilities of drum onsets, beats, and downbeats are estimated in a conditionally independent manner and are thus not guaranteed to be consistent in terms of periodicity.

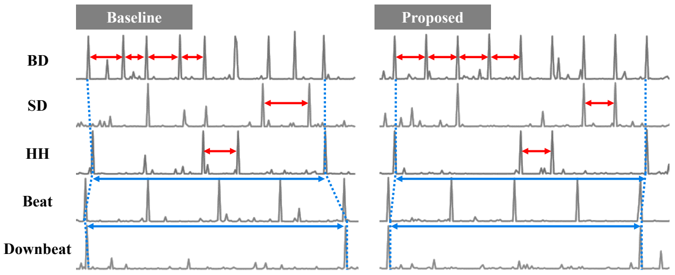

Fig. 1. The frame-level sequences of drum onset, beat, and downbeat probabilities estimated by the multi-task learning method without/with the periodicity-aware regularization. The red and blue arrows indicate the element-wise periodicity and pairwise consistency, respectively.

To mitigate this problem, we propose a regularized training method that encourages the sequences of drum onset, beat, and downbeat probabilities to have periodic patterns synchronously (Fig. 1). We focus on the fact that the tempo tends to be kept constant throughout the entire song in popular music. This requests that each sequence should have clear periodicity solely and that a sequence should share the same periodicity with another sequence. We thus propose element-wise and pairwise regularization terms described in detail below.

First, we focus on the auto-correlation function of each sequence. If the sequence of drum onset probabilities is repetitive at the measure level, its auto-correlation function exhibits sharp peaks spaced at the downbeat interval. The auto-spectrum obtained by applying the discrete Fourier transform (DFT) to the auto-correlation function exhibits a few peaks at the downbeat frequency. The *entropy of the auto-spectrum* can thus be used as a regularization term to be minimized. The same applies to the sequences of beat and downbeat probabilities exhibiting the periodicity at the beat and downbeat levels.

Second, we focus on the cross-correlation function for each of the three sequence pairs. In general, every time the sequence

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of drum onset probabilities is time-shifted by a downbeat interval, it matches well the sequence of downbeat probabilities in terms of peak positions. The cross-correlation function exhibits sharp peaks spaced at the downbeat interval. The *entropy of the cross-spectrum* can thus be another regularization term. The same applies to the other two pairs.

We experimentally show that the multi-task learning approach is especially effective for improving the performance of metrical analysis and the regularized training method has a potential of improving the performances of drum transcription and metrical analysis.

II. Related Work

Deep learning has been the central choice in modern ADT [2], [3], [5]–[10]. In general, a deep neural network (DNN) is trained in a supervised manner using paired data consisting of music signals and drum onset annotations. Due to the impulsive nature of drums sounds, CNNs have widely been used as basic building blocks for extracting local features. In particular, a CNN variant called the temporal convolutional network (TCN) [11], [12] has gained much attention because it can take into account both short- and long-term dependencies [13]. To consider long-term contextual information, a CNN is often followed by a recurrent neural network (RNN), resulting in a convolutional RNN (CRNN) [2].

To improve the performance with a limited amount of paired data, one may use data synthesis [14], data augmentation [5], or unsupervised learning [15]. However, these methods are still insufficient for dealing with considerable variations in real-world drum sounds. Alternatively, regularized training techniques based on prior knowledge of drum patterns have been explored [10] as well as phase consideration [16] and architectural innovations [7]. In this study, we focus on the periodicity of drum onsets for regularization.

For metrical analysis, DNN-HMM combinations have often been used [13], [17]–[20]. In this approach, a DNN is used for estimating sequences of beat and downbeat probabilities, and an HMM is then used for detecting beat and downbeat times from these sequences. Instead of the standard HMM, a hidden semi-Markov model (HSMM) that explicitly represents the duration of each state was proven to be effective for taking into account the tempo and meter consistency of beat and downbeat times. In this study, we use a similar HSMM-based decoder to obtain consistent results.

III. Proposed Method

This section describes the multi-task learning of drum transcription and metrical analysis with a periodicity-aware regularized training method.

A. Problem Specification

We tackle drum transcription and beat and downbeat detection from a music signal. We use the power spectrograms of the left and right channels of the audio signal $\mathbf{X} \in \mathbb{R}^{2 \times F \times T}$ as input. Let $\mathbf{Y}^{\mathrm{D}} \in \{0,1\}^{K \times T}$, $\mathbf{Y}^{\mathrm{B}} \in \{0,1\}^{T}$, and $\mathbf{Y}^{\mathrm{W}} \in \{0,1\}^{T}$ denote the presence or absence of drum onsets, that of beats, and that of downbeats at the frame level, respectively, where F is the number of frequency bins, T is the number of frames, and K is the number of drum instruments. We focus on the three major drum classes ($K = 3$): bass drum (BD), snare drum (SD), and hi-hats (HH). Let $\mathbf{Y} \triangleq \{\mathbf{Y}^{\mathrm{D}}, \mathbf{Y}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{W}}\}$. In the training phase, both $\mathbf{X}$ and $\mathbf{Y}$ are given as paired data. In the test phase, $\mathbf{Y}$ is estimated from $\mathbf{X}$.

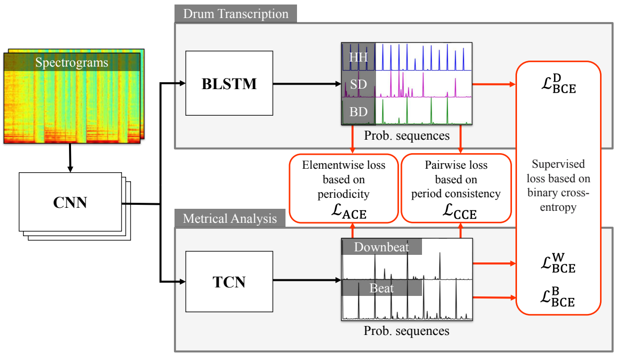
Drum Transcription
Spectrograms
BLSTM
HH
SD
BD
Prob. sequences
$\mathcal{L}^{\mathrm{D}}_{\mathrm{BCE}}$
CNN
Elementwise loss based on periodicity $\mathcal{L}_{\mathrm{ACE}}$
Pairwise loss based on period consistency $\mathcal{L}_{\mathrm{CCE}}$
Supervised loss based on binary cross-entropy
Metrical Analysis
TCN
Downbeat
Beat
Prob. sequences
$\mathcal{L}^{\mathrm{W}}_{\mathrm{BCE}}$
$\mathcal{L}^{\mathrm{B}}_{\mathrm{BCE}}$

Fig. 2. The proposed joint drum transcription and metrical analysis based on periodicity-aware regularized multi-task learning.

B. Multi-Task Learning

We explain the loss function to be minimized for joint drum transcription and metrical analysis. Inspired by [17], [20], we use a CRNN with a branching architecture for jointly estimating a sequence of drum onset probabilities $\boldsymbol{\phi}^{\mathrm{D}} \in [0,1]^{K \times T}$, that of beat probabilities $\boldsymbol{\phi}^{\mathrm{B}} \in [0,1]^{T}$, and that of downbeat probabilities $\boldsymbol{\phi}^{\mathrm{D}} \in [0,1]^{T}$ from the input $\mathbf{X}$ (Fig. 2).

1) Drum Transcription Loss: The drum transcription loss is given by the binary cross-entropy (BCE) of $\boldsymbol{\phi}^{\mathrm{D}}$ for the ground-truth data $\mathbf{Y}^{\mathrm{D}}$:

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{BCE}} = -\frac{1}{K} \sum_{k,t=1}^{K,T} \left(Y^{\mathrm{D}}_{kt} \log \phi^{\mathrm{D}}_{kt} + (1 - Y^{\mathrm{D}}_{kt}) \log(1 - \phi^{\mathrm{D}}_{kt}) \right). \quad (1)$$

2) Metrical Analysis Losses: The metrical analysis losses are given by the BCEs of $\boldsymbol{\phi}^{\mathrm{B}}$ and $\boldsymbol{\phi}^{\mathrm{W}}$ for the ground-truth data $\mathbf{Y}^{\mathrm{B}}$ and $\mathbf{Y}^{\mathrm{W}}$:

$$\mathcal{L}^{\mathrm{B}}_{\mathrm{BCE}} = -\sum_{t=1}^{T} \left(Y^{\mathrm{B}}_{t} \log \phi^{\mathrm{B}}_{t} + (1 - Y^{\mathrm{B}}_{t}) \log(1 - \phi^{\mathrm{B}}_{t}) \right), \quad (2)$$

$$\mathcal{L}^{\mathrm{W}}_{\mathrm{BCE}} = -\sum_{t=1}^{T} \left(Y^{\mathrm{W}}_{t} \log \phi^{\mathrm{W}}_{t} + (1 - Y^{\mathrm{W}}_{t}) \log(1 - \phi^{\mathrm{W}}_{t}) \right). \quad (3)$$

3) Total Loss: For joint training, we aim to minimize the weighted sum of the individual losses:

$$\mathcal{L}_{\mathrm{BCE}} = \lambda^{\mathrm{D}}_{\mathrm{BCE}} \mathcal{L}^{\mathrm{D}}_{\mathrm{BCE}} + \lambda^{\mathrm{B}}_{\mathrm{BCE}} \mathcal{L}^{\mathrm{B}}_{\mathrm{BCE}} + \lambda^{\mathrm{W}}_{\mathrm{BCE}} \mathcal{L}^{\mathrm{W}}_{\mathrm{BCE}}, \quad (4)$$

where $\lambda^{\mathrm{D}}_{\mathrm{BCE}}$, $\lambda^{\mathrm{B}}_{\mathrm{BCE}}$, and $\lambda^{\mathrm{W}}_{\mathrm{BCE}}$ are adjustable weights.

C. Periodicity-Aware Regularization

We explain regularization terms that encourage elementwise periodicity and pairwise consistency. We first explain the core ideas underlying the regularization terms (Figs. 3(a) and 3(b)) and describe the regularization terms.

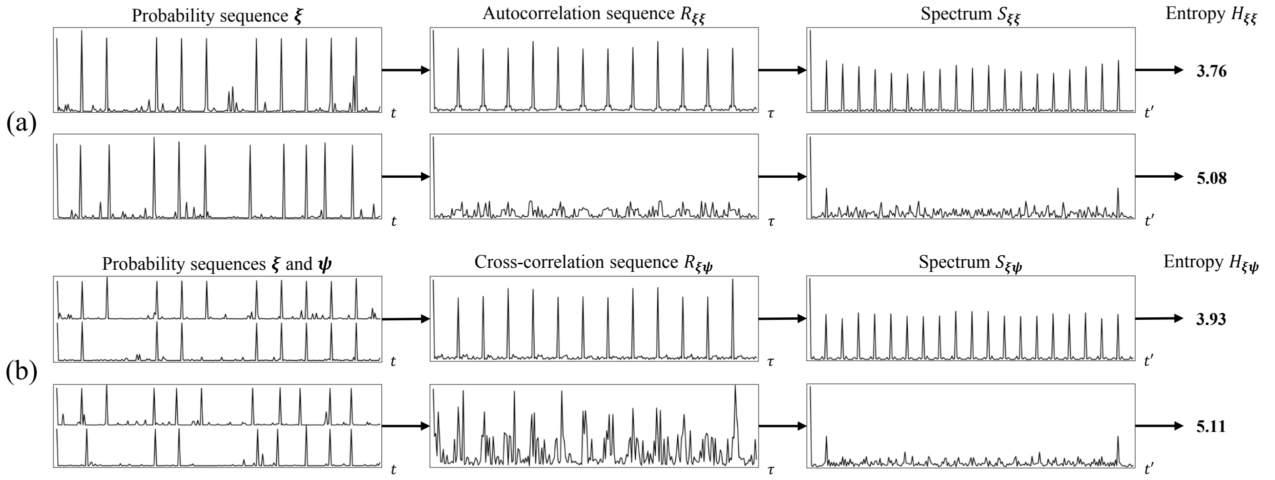

Fig. 3. (a) Examples of probability sequences and the corresponding auto-correlation functions and auto-spectra with entropies. A more periodic probability sequence results in the auto-spectrum with a smaller entropy. (b) Examples of probability sequence pairs and the corresponding cross-correlation functions and cross-spectra with entropies. A more consistent probability sequence pair results in the cross-spectrum with a smaller entropy.

1) Auto-Correlation Analysis: The auto-correlation function for a probability sequence $\boldsymbol{\xi} \in [0,1]^T$ is given by

$$R_{\boldsymbol{\xi}\boldsymbol{\xi}}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \xi_t \xi_{(t+\tau-1)\%T+1}, \tag{5}$$

where $\%$ represents the modulo operation. Using the Wiener-Khinchin theorem, the spectrum of $R_{\boldsymbol{\xi}\boldsymbol{\xi}}$ is given by

$$S_{\boldsymbol{\xi}\boldsymbol{\xi}} = FR_{\boldsymbol{\xi}\boldsymbol{\xi}} = |F\boldsymbol{\xi}|^{\cdot 2}, \tag{6}$$

where $F \in \mathbb{C}^{T\times T}$ represents the discrete Fourier transform (DFT) matrix and $|\mathbf{a}|^{\cdot 2}$ represents the element-wise operation that takes the absolute square values of the elements of a vector $\mathbf{a}$. The entropy of the spectrum $S_{\boldsymbol{\xi}\boldsymbol{\xi}}$ is given by

$$H_{\boldsymbol{\xi}\boldsymbol{\xi}} = -\sum_{t=1}^{T} \bar{S}_{\boldsymbol{\xi}\boldsymbol{\xi}}(t)\log \bar{S}_{\boldsymbol{\xi}\boldsymbol{\xi}}(t), \tag{7}$$

$$\bar{S}_{\boldsymbol{\xi}\boldsymbol{\xi}} = \mathrm{Normalize}(S_{\boldsymbol{\xi}\boldsymbol{\xi}}), \tag{8}$$

where $\mathrm{Normalize}(\mathbf{a})$ represents the normalization operation that makes the sum of the elements of a vector $\mathbf{a}$ equal to 1. If the probability sequence $\boldsymbol{\xi}$ has clear periodicity, i.e., it has regular periodic patterns and the peak intervals are integer multiples of some basic time unit (e.g., tatum, beat, or downbeat), the auto-correlation function and the auto-spectrum have equally-spaced sharp peaks. The negative entropy $-H_{\boldsymbol{\xi}\boldsymbol{\xi}}$ can be used as an indicator of periodicity.

2) Cross-Correlation Analysis: The cross-correlation function for probability sequences $\boldsymbol{\xi}, \boldsymbol{\psi} \in [0,1]^T$ is given by

$$R_{\boldsymbol{\xi}\boldsymbol{\psi}}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \xi_t \psi_{(t+\tau-1)\%T+1}. \tag{9}$$

Using the Wiener-Khinchin theorem, the magnitude spectrum of $R_{\boldsymbol{\xi}\boldsymbol{\psi}}$ is given by

$$S_{\boldsymbol{\xi}\boldsymbol{\psi}} = |FR_{\boldsymbol{\xi}\boldsymbol{\psi}}|^{\cdot} = |(F\boldsymbol{\xi})^* \odot (F\boldsymbol{\psi})|^{\cdot}, \tag{10}$$

where $\mathbf{a}^*$ represents the element-wise conjugate operation for a vector $\mathbf{a}$, $\odot$ represents the element-wise multiplication, and $|\mathbf{a}|^{\cdot}$ represents the element-wise operation that takes the absolute values of the elements of a vector $\mathbf{a}$. The entropy of the spectrum $S_{\boldsymbol{\xi}\boldsymbol{\psi}}$ is given by

$$H_{\boldsymbol{\xi}\boldsymbol{\xi}} = -\sum_{t=1}^{T} \bar{S}_{\boldsymbol{\xi}\boldsymbol{\psi}}(t)\log \bar{S}_{\boldsymbol{\xi}\boldsymbol{\psi}}(t), \tag{11}$$

$$\bar{S}_{\boldsymbol{\xi}\boldsymbol{\psi}} = \mathrm{Normalize}(S_{\boldsymbol{\xi}\boldsymbol{\psi}}). \tag{12}$$

If the probability sequences $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$ are consistent in terms of periodicity, i.e., they have synchronous periodic patterns, the cross-correlation function and the cross-spectrum have equally-spaced sharp peaks. The negative entropy $-H_{\boldsymbol{\xi}\boldsymbol{\psi}}$ can be used as an indicator of consistency.

3) Element-Wise Regularization: We propose element-wise regularization terms, denoted by $\mathcal{L}^{\mathrm{D}}_{\mathrm{ACE}}$, $\mathcal{L}^{\mathrm{B}}_{\mathrm{ACE}}$, and $\mathcal{L}^{\mathrm{W}}_{\mathrm{ACE}}$, based on the auto-correlation entropy (ACE) of the drum onset probability sequence $\boldsymbol{\phi}^{\mathrm{D}}$, that of the beat probability sequence $\boldsymbol{\phi}^{\mathrm{B}}$, and that of the downbeat probability sequence $\boldsymbol{\phi}^{\mathrm{W}}$:

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{ACE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T} \bar{\phi}^{\mathrm{D}}_{kt}\log\bar{\phi}^{\mathrm{D}}_{kt}, \tag{13}$$

$$\bar{\boldsymbol{\phi}}^{\mathrm{D}}_{k} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{D}}_{k}|^{\cdot 2}), \tag{14}$$

$$\mathcal{L}^{\mathrm{B}}_{\mathrm{ACE}} = -\sum_{t=1}^{T} \bar{\phi}^{\mathrm{B}}_{t}\log\bar{\phi}^{\mathrm{B}}_{t}, \tag{15}$$

$$\bar{\boldsymbol{\phi}}^{\mathrm{B}} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{B}}|^{\cdot 2}), \tag{16}$$

$$\mathcal{L}^{\mathrm{W}}_{\mathrm{ACE}} = -\sum_{t=1}^{T} \bar{\phi}^{\mathrm{W}}_{t}\log\bar{\phi}^{\mathrm{W}}_{t}, \tag{17}$$

$$\bar{\boldsymbol{\phi}}^{\mathrm{W}} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{W}}|^{\cdot 2}), \tag{18}$$

where $\boldsymbol{\phi}^{\mathrm{D}}_{k} \in [0,1]^T$ denotes the drum onset probability sequence for drum class k.

We aim to minimize the weighted sum of the individual regularization terms:

$$\mathcal{L}_{\text{ACE}} = \lambda^{\text{D}}_{\text{ACE}}\mathcal{L}^{\text{D}}_{\text{ACE}} + \lambda^{\text{B}}_{\text{ACE}}\mathcal{L}^{\text{B}}_{\text{ACE}} + \lambda^{\text{W}}_{\text{ACE}}\mathcal{L}^{\text{W}}_{\text{ACE}}, \tag{19}$$

where $\lambda^{\text{D}}_{\text{ACE}}$, $\lambda^{\text{B}}_{\text{ACE}}$, and $\lambda^{\text{W}}_{\text{ACE}}$ are adjustable weights.

4) Pairwise Regularization: We propose pairwise regularization terms, denoted by $\mathcal{L}^{\text{DB}}_{\text{CCE}}$, $\mathcal{L}^{\text{DW}}_{\text{CCE}}$, and $\mathcal{L}^{\text{BW}}_{\text{CCE}}$, based on the cross-correlation entropy (CCE) for each of the possible pairs between the drum onset probability sequence $\boldsymbol{\phi}^{\text{D}}$, the beat probability sequence $\boldsymbol{\phi}^{\text{B}}$, and the downbeat probability sequence $\boldsymbol{\phi}^{\text{W}}$:

$$\mathcal{L}^{\text{DB}}_{\text{CCE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T} \bar{\phi}^{\text{DB}}_{kt} \log \bar{\phi}^{\text{DB}}_{kt}, \tag{20}$$

$$\bar{\boldsymbol{\phi}}^{\text{DB}}_{k} = \text{Normalize}(|F\boldsymbol{\phi}^{\text{D}}_{k}|^{\cdot} \odot |F\boldsymbol{\phi}^{\text{B}}|^{\cdot}), \tag{21}$$

$$\mathcal{L}^{\text{DW}}_{\text{CCE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T} \bar{\phi}^{\text{DW}}_{kt} \log \bar{\phi}^{\text{DW}}_{kt}, \tag{22}$$

$$\bar{\boldsymbol{\phi}}^{\text{DW}}_{k} = \text{Normalize}(|F\boldsymbol{\phi}^{\text{D}}_{k}|^{\cdot} \odot |F\boldsymbol{\phi}^{\text{W}}|^{\cdot}), \tag{23}$$

$$\mathcal{L}^{\text{BW}}_{\text{CCE}} = -\sum_{t=1}^{T} \bar{\phi}^{\text{BW}}_{t} \log \bar{\phi}^{\text{BW}}_{t}, \tag{24}$$

$$\bar{\boldsymbol{\phi}}^{\text{BW}} = \text{Normalize}(|F\boldsymbol{\phi}^{\text{B}}|^{\cdot} \odot |F\boldsymbol{\phi}^{\text{W}}|^{\cdot}). \tag{25}$$

We aim to minimize the weighted sum of the individual regularization terms:

$$\mathcal{L}_{\text{CCE}} = \lambda^{\text{DB}}_{\text{CCE}}\mathcal{L}^{\text{DB}}_{\text{CCE}} + \lambda^{\text{DW}}_{\text{CCE}}\mathcal{L}^{\text{DW}}_{\text{CCE}} + \lambda^{\text{BW}}_{\text{CCE}}\mathcal{L}^{\text{BW}}_{\text{CCE}}, \tag{26}$$

where $\lambda^{\text{DB}}_{\text{CCE}}$, $\lambda^{\text{DW}}_{\text{CCE}}$, and $\lambda^{\text{BW}}_{\text{CCE}}$ are adjustable weights.

5) Total Loss: For regularized joint training, we aim to minimize the sum of the basic supervised loss $\mathcal{L}_{\text{BCE}}$ in (4), the element-wise regularization term $\mathcal{L}_{\text{ACE}}$ in (19), and the pairwise regularization term $\mathcal{L}_{\text{CCE}}$ in (26):

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{ACE}} + \mathcal{L}_{\text{CCE}}. \tag{27}$$

The weights can be adjusted for controlling the relative importance of different tasks.

D. Decoding

In the test phase, given $\mathbf{X}$, we jointly estimate the drum onset probability sequence $\boldsymbol{\phi}^{\text{D}}$, the beat probability sequence $\boldsymbol{\phi}^{\text{B}}$, and the downbeat probability sequence $\boldsymbol{\phi}^{\text{W}}$ using the trained CRNN. For improved performance, we use different decoding techniques for detecting the drum onsets $\mathbf{Y}^{\text{D}}$ and the beat and downbeat times $\mathbf{Y}^{\text{D}}$ and $\mathbf{Y}^{\text{W}}$.

1) Drum Transcription: The final output $\mathbf{Y}^{\text{D}}$ is obtained by performing peak-picking and thresholding for $\boldsymbol{\phi}^{\text{D}}$ as follows:

$$Y^{\text{D}}_{kt} = \begin{cases} 1 & (\phi^{\text{D}}_{kt} \geq \delta_k) \\ 0 & (\text{otherwise}) \end{cases}, \tag{28}$$

where δ_k is a threshold configured for each drum instrument.

2) Metrical Analysis: Since the naive thresholding has a limitation in performance, the final outputs $\mathbf{Y}^{\text{B}}$ and $\mathbf{Y}^{\text{W}}$ are detected from from $\boldsymbol{\phi}^{\text{B}}$ and $\boldsymbol{\phi}^{\text{W}}$ with a metrical analysis method based on a dynamic Bayesian network (DBN) [21].

We explain the decoding method for the beat probability sequence $\boldsymbol{\phi}^{\text{B}}$. The DBN is a hidden Markov model (HMM) with hidden states $\mathbf{z}_t = [c_t, \dot{c}_t]$ at frame t. Here, $c_t \in \{1, 2, \ldots, \dot{c}\}$ is a discrete random variable representing the position within a beat at frame t, and $\dot{c}_t \in \{\dot{c}_{\text{min}}, \dot{c}_{\text{min}} + 1, \ldots, \dot{c}_{\text{max}}\}$ is the total number of discretized positions within one beat, which is a random variable representing the tempo at frame t, where $\dot{c}_{\text{min}}$ and $\dot{c}_{\text{max}}$ denote the minimum and maximum values of the tempo, respectively. Let Δ be the frame length of the audio signal and BPM_{t} represent the tempo in beats per minute at frame t. Then, $\dot{c}_t$ can be expressed as follows:

$$\dot{c}_t = \text{round}\left(\frac{4 \times 60}{\text{BPM}_{\text{t}} * \Delta}\right). \tag{29}$$

If we denote the observation sequence as $\{\mathbf{o}_t\}_{t=1}^{T}$, the probability model can be defined as follows:

$$p(\mathbf{z}_{1:T}, \mathbf{o}_{1:T}) = p(\mathbf{z}_{1:T})p(\mathbf{o}_{1:T}|\mathbf{z}_{1:T}). \tag{30}$$

Our goal is to estimate the hidden state sequence $\mathbf{z}^{*}_{1:T}$ with the maximum posterior probability as follows:

$$\mathbf{z}^{*}_{1:T} = \underset{z_{1:T}}{\text{argmax}}\, p(\mathbf{z}_{1:T}|\mathbf{o}_{1:T}), \tag{31}$$

$$p(\mathbf{o}_{1:T}|\mathbf{z}_{1:T}) \propto p(\mathbf{z}_1)\prod_{t=2}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{o}_t|\mathbf{z}_t), \tag{32}$$

where $p(\mathbf{o}_t|\mathbf{z}_t)$ represents the observation model, and $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ represents the transition model. Equation (31) can be solved efficiently using the Viterbi algorithm [22]. The term $p(\mathbf{z}_{1:T})$ in (30) represents the generative process of the hidden states $\mathbf{z}_{1:T}$, which is given by

$$p(\mathbf{z}_{1:T}) = p(\mathbf{z}_1)\prod_{t=2}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1}), \tag{33}$$

where $p(\mathbf{z}_1)$ represents the initial probability and $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ represents the transition probability given by

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) = p(c_t|c_{t-1}, \dot{c}_{t-1})p(\dot{c}_t|\dot{c}_{t-1}), \tag{34}$$

$$p(c_t|c_{t-1}, \dot{c}_{t-1}) = \mathbb{1}_z. \tag{35}$$

where $\mathbb{1}_z$ is a function that takes 1 if $c_t = (c_{t-1}+1) \bmod \dot{\text{c}}_{t-1}$ and 0 otherwise. If $c_{t-1} = \dot{c}_{t-1}$, we have

$$p(\dot{c}_t|\dot{c}_{t-1}) = \exp\left(-\lambda\left|\frac{\dot{c}_t}{\dot{c}_{t-1}} - 1\right|\right), \tag{36}$$

where λ is a hyperparameter. Otherwise, we have

$$p(\dot{c}_t|\dot{c}_{t-1}) = \begin{cases} 1 & (c_t = c_{t-1} + 1) \\ 0 & \text{otherwise} \end{cases}. \tag{37}$$

Finally, for the hidden state sequence $\mathbf{x}^{*}_{1:T}$ with the maximum posterior probability, we detect all t's where $c_t = 1$. This sequence is the estimated beat time sequence.

TABLE I
THE F-MEASURES OF DRUM TRANSCRIPTION AND METRICAL ANALYSIS. VALUES IN BOLD ARE WITHIN 0.5 PTS OF THE BEST VALUE.

Multi-task Learning			Regularized Training		Drum Transcription				Metrical Analysis	
D	B	W	ACE	CCE	BD	SD	HH	Ave.	B	W
	✓	✓			-	-	-	-	90.5	87.3
✓					**78.0**	71.2	74.2	74.5	-	-
✓			✓		77.0	70.1	74.7	73.9	-	-
✓	✓				76.6	69.6	76.5	74.2	90.6	-
✓	✓		✓		76.6	71.2	75.2	74.3	90.3	-
✓	✓			✓	77.0	71.3	75.1	74.5	88.9	-
✓	✓		✓	✓	**77.6**	**73.5**	75.7	**75.6**	89.5	-
✓	✓	✓			75.1	68.7	72.4	72.1	93.9	**89.2**
✓	✓	✓	✓		77.1	74.1	73.7	75.0	**94.6**	**89.1**
✓	✓	✓		✓	74.2	70.7	76.4	73.8	**94.1**	87.2
✓	✓	✓	✓	✓	76.8	72.2	**77.1**	**75.4**	92.1	86.3

IV. EVALUATION

This section reports a comparative experiment conducted for evaluating the proposed regularized multi-task learning method through ablation study.

A. Experimental Conditions

The RWC Music Database: Popular Music [23] was used for evaluation. It consists of 100 Japanese popular songs (J-POP) with human-performed and synthesized drum tracks. We used 64 songs with accurate drum onset annotations and 10 songs without drum annotations. The stereo signals at a sampling rate of 44.1 kHz were processed using short-time Fourier transform (STFT) with a window size of 1024 points and a hop size of 441 points. The left and right channels were concatenated to form the input matrix $\mathbf{X}$.

Our CRNN consists of common convolutional layers for feature extraction followed by a bidirectional long short-term memory (BLSTM) network for drum transcription and a TCN for metrical analysis (Fig. 2). The convolutional layers, with a kernel size of 3×3, a padding size of 1×1, and a stride of 1, yielded a (512×4)-dimensional feature map, on which a dropout of 30% was applied before feeding it into a linear layer. The BLSTM network had three layers, each of which had hidden states of 200 dimensions. The TCN had eleven layers with a receptive field of about 80 [s]. The weights were set as follows: $\lambda^{\mathrm{D}}_{\mathrm{BCE}}{=}1$, $\lambda^{\mathrm{B}}_{\mathrm{BCE}}{=}\lambda^{\mathrm{W}}_{\mathrm{BCE}}{=}0.1$, $\lambda^{\mathrm{D}}_{\mathrm{ACE}}{=}1$, $\lambda^{\mathrm{B}}_{\mathrm{ACE}}{=}\lambda^{\mathrm{W}}_{\mathrm{ACE}}{=}0$, and $\lambda^{\mathrm{DB}}_{\mathrm{CCE}}{=}\lambda^{\mathrm{DW}}_{\mathrm{CCE}}{=}\lambda^{\mathrm{BW}}_{\mathrm{CCE}}{=}1$. We used AdamW [24] with a learning rate of $\gamma = 0.001$, weight decays parameters of $\lambda = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\varepsilon = 10^{-9}$ for optimization.

To evaluate the effectiveness of the individual components of the proposed method, we conducted an ablation study. We tackled drum transcription and metrical analysis jointly or separately with or without the element-wise and pairwise regularization terms. Five-fold cross-validation was conducted, where the performance was evaluated on the 64 songs with drum annotations. We also compared the proposed method with a state-of-the-art method [20].

The performances of drum transcription and metrical analysis were evaluated in terms of the F-measure with error tolerances of 30 [ms] and 70 [ms], respectively. The beat interval stability was also evaluated in terms of the CMLt and AMLt metrics [25] that consider double/half tempo ambiguity in beat interpretation. Note that the drum onset annotations might be inconsistent due to the characteristics of annotators. In the training phase, the ground-truth onset times were perturbed by being convolved with a Gaussian distribution with a mean of 0 [ms] and a standard deviation of 12 [ms]. For data augmentation, the time stretch operation was applied to each song with a scaling factor drawn from a Gaussian distribution with a mean of 1 and a standard deviation of 0.1.

TABLE II
THE PERFORMANCES OF METRICAL ANALYSIS FOR 100 SONGS OF RWC POPULAR MUSIC DATABASE. BOLD FONTS INDICATE THE BEST VALUES.

	Beat			Downbeat		
	F-meas.	CMLt	AMLt	F-meas.	CMLt	AMLt
Böck et al. [20]	89.5	81.8	91.5	83.1	80.3	88.4
Ours	**92.8**	**85.9**	**94.3**	**86.3**	**81.3**	**89.4**

B. Experimental Results

As shown in Table I, the multi-task learning approach improved the beat and downbeat detection performances by approximately 4 and 2 pts, respectively, compared with the single-task learning approach. Both the elementwise and pairwise regularizations improved the performance of drum transcription by approximately 1–3 pts. For metrical analysis, in contrast, no significant improvement was observed, mainly because the DBN-based decoding had a dominant impact on the results. If the naive thresholding was used instead, some performance gain was obtained (not reported in the table). As shown in Table II, the configuration that achieved the best performance of beat detection for all the 100 songs of the RWC Popular Music Database (the ninth row in Table I) outperformed the state-of-the-art method [20] by approximately 3 pts higher.

Fig. 4 shows the positive and negative effects of the multi-task learning. In the left example, since the drum onsets were located on the beat and tatum grids, the snare drum onsets on unnatural positions were suppressed successfully. In contrast, in the right example, although the drum onsets with swing rhythm were deviated from the beat and tatum grids, they were encouraged to be synchronized with the grids.

Fig. 5 shows the positive effect of the regularized training. The histogram of the inter-onset intervals (IOIs) of the hi-hat onsets had a single sharp peak, meaning that the estimated onsets were placed at a constant interval. Such periodicity-aware regularization, however, might fail to estimate non-regular drum patterns (e.g., fill-ins) and degrade the performance.

These results suggest the importance of the weight configuration according to the target data. One promising solution worth investigation would be to use the self-attention mechanism for automatically adjusting the weights according to the regularity of drum patterns.

V. CONCLUSION

In this paper, we presented a multi-task learning approach to joint drum transcription and metrical analysis based on the periodicity-aware element-wise and pairwise regularization terms. Specifically, our method aims to minimize the entropy

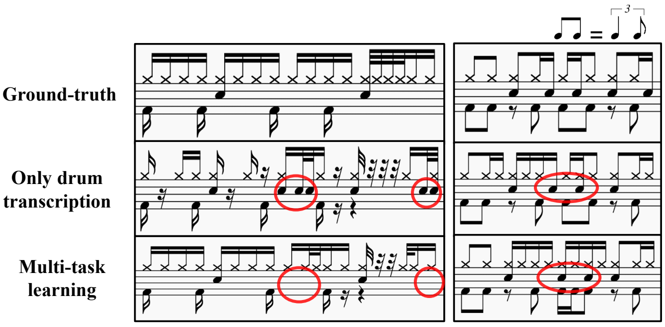

Fig. 4. Examples of drum transcription results obtained with the multi-task and single-task learning approaches. The performance was improved in the left example and was degraded in the right example.

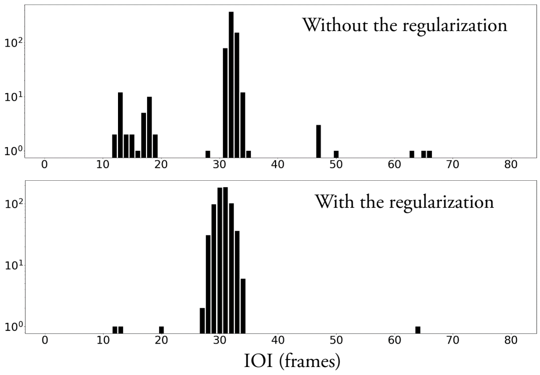

Fig. 5. Histograms of inter-onset intervals (IOIs) in frame units computed from hi-hat onsets detected without/with the regularization (RWC-MDB-P-2001 No. 11).

of the auto-spectrum computed from each probability sequence and that of the cross-spectrum computed from each of the three sequence pairs, because a smaller entropy indicates a stronger element-wise periodicity or pairwise consistency. We experimentally showed the effectiveness of using the multi-task learning and the regularized training for drum transcription. The performance of metrical analysis, however, was not maximized by the concurrent use of these methods. Further investigation is thus needed to determine the weights of the loss functions appropriately for each task, compare with other deep learning methods, and explore a peak detection method that maximizes the performance of the proposed method. In addition, we plan to investigate a new regularization term based on the fractal structure inherent in music.

Acknowledgment

This study was partially supported by JSPS KAKENHI Nos. 21K12187, 21K02846, 22H03661, 20H00602, and 21H03572, JST PRESTO No. JPMJPR20CB, and JST FOREST No. JPMJPR226X.

References

[1] C. W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Müller, and A. Lerch. A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 9, pp. 1457–1483, 2018.

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

[3] C. Jacques and A. Roebel. Automatic drum transcription with convolutional neural networks. In *International Conference on Digital Audio Effects (DAFx)*, pp. 80–86, 2018.

[4] E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *The Journal of the Acoustical Society of America*, Vol. 133, No. 3, pp. 1727–1741, 2013.

[5] C. Jacques and A. Roebel. Data augmentation for drum transcription with convolutional neural networks. In *European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

[6] R. Stables, J. Hockman, and C. Southall. Automatic drum transcription using bi-directional recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 591–597, 2016.

[7] S. Ueda, K. Shibata, Y. Wada, R. Nishikimi, E. Nakamura, and K. Yoshii. Drum transcription using convolutional non-negative matrix factorization based on deep drum score prior distribution (in japanese). *IPSJ SIG technical reports (EC)*, Vol. 2019, No. 26, pp. 1–6, 2019.

[8] R. Ishizuka, R. Nishikimi, E. Nakamura, and K. Yoshii. Tatum-level drum transcription based on a convolutional recurrent neural network with language model-based regularized training. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020.

[9] Y. Wang, J. Salamon, M. Cartwright, Nicholas J. Bryan, and J. P. Bello. Few-shot drum transcription in polyphonic music. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 117–124, 2020.

[10] R. Ishizuka, R. Nishikimi, and K. Yoshii. Global structure-aware drum transcription based on self-attention mechanisms. *Signals*, Vol. 2, No. 3, pp. 508–526, 2021.

[11] C. Lea, R. Vidal, A. Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision (ECCV) Workshop*, pp. 47–54, 2016.

[12] S. Bai, J. Zico Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In *arXiv preprint arXiv:1803.01271*, 2018.

[13] M. E. P. Daveis and S. Böck. Temporal convolutional networks for musical audio beat tracking. In *European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

[14] M. Cartwright and J. P. Bello. Increasing drum transcription vocabulary using data synthesis. In *International Conference on Digital Audio Effects (DAFx)*, pp. 72–79, 2018.

[15] K. Choi and K. Cho. Deep unsupervised drum transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 183–191, 2019.

[16] H. Kawata, T. Hori, and K. Nakamura. Automatic drum transcription based on RNN considering phase information (in japanese). *IPSJ SIG technical reports (MUS)*, No. 27, pp. 1–4, 2019.

[17] T. Oyama, R. Ishizuka, and K. Yoshii. Phase-aware joint beat and downbeat estimation based on periodicity of metrical structure. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–499, 2021.

[18] S. Böck, F. Krebs, and G. Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 255–261, 2016.

[19] S. Böck, M. E. P. Davies, and P. Knees. Multi-task learning of tempo and beat: Learning one to improve the other. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 486–493, 2019.

[20] S. Böck and M. E. P. Daveis. Deconstruct, analysis, reconstruct: How to improve tempo, beat, and downbeat estimation. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 574–582, 2020.

[21] F. Krebs, S. Böck, and G. Widmer. An efficient state-space model for joint tempo and meter tracking. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 72–78, 2015.

[22] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE.*, pp. 77(2):257–286, 1989.

[23] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, classical and jazz music databases. In *International*

Society for Music Information Retrieval Conference (ISMIR), pp. 287–288, 2002.

[24] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, pp. 1–8, 2017.

[25] M. E. P. Davies, N. Degara, and M. D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. Technical report, Centre for Digital Music, Queen Mary University of London, 2009.