Tatum-Level Drum Transcription Based on a Convolutional Recurrent Neural Network with Language Model-Based **Regularized Training**

Graduate School of Informatics, Kyoto University, Kyoto, Japan

Ryoto Ishizuka, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii

Overview

Overview

CHAPTER 1

Background

Automatic music transcription (AMT)

Aim : Estimate music scores from audio signals

Value : Help music composition and arrangement

Automatic drum transcription (ADT)

Role : Rhythmic backbone of popular music

Inst : Multiple

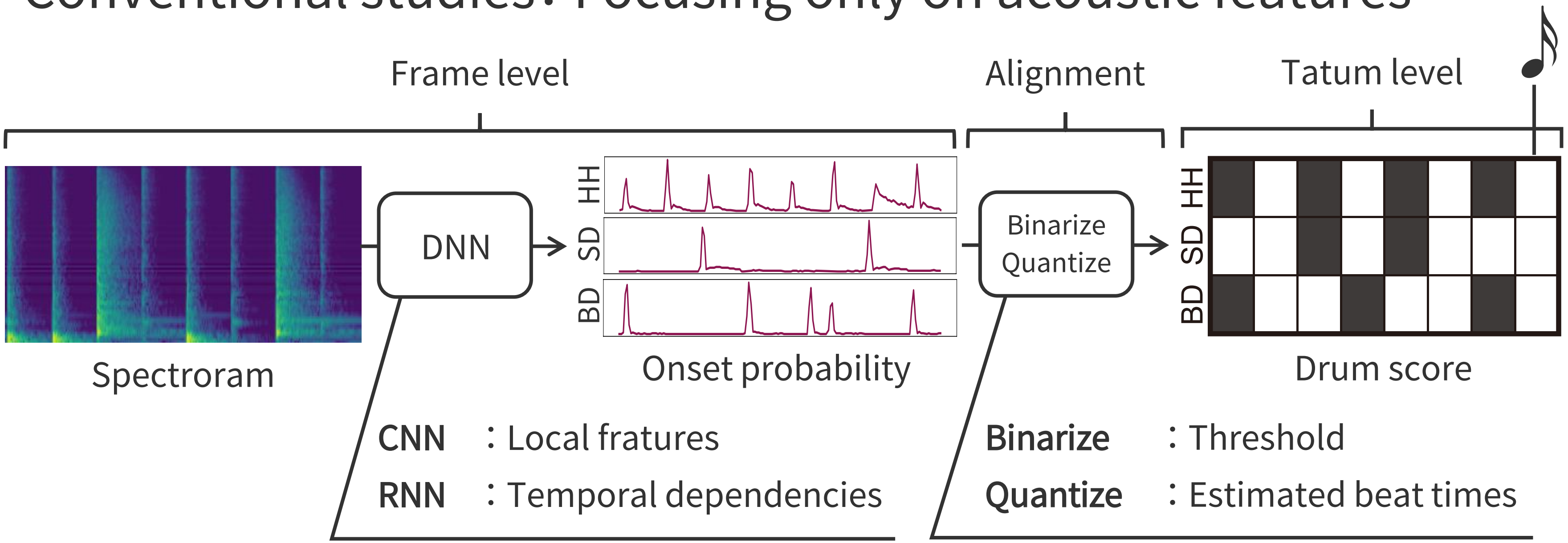
Pitch : Different from instruments

Value : Hard to adjust

Most works focus on onset times which the main three parts are played at

Background

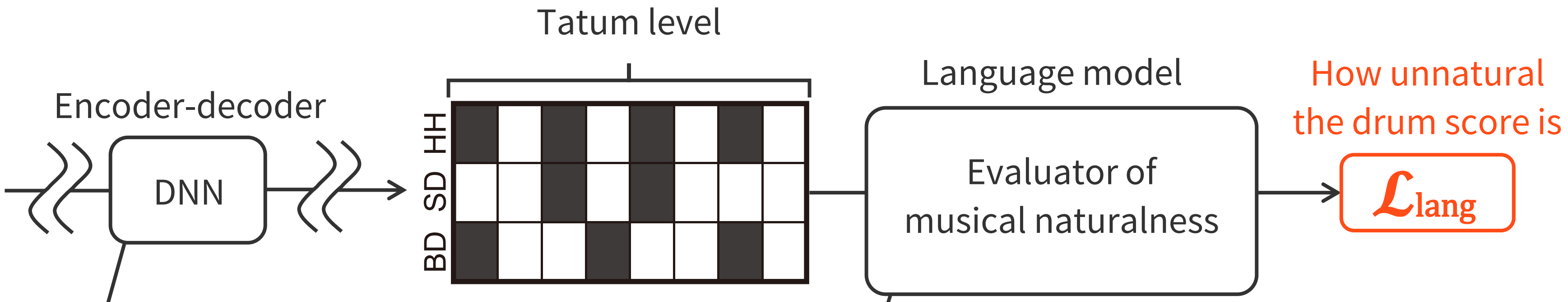
Conventional studies: Focusing only on acoustic features



Problem☹: Often estimates musically-unnatural drum patterns

Background

Idea：Tatum-level **language model-based** ADT method



1. CNN-based encoder
2. GRU-based decoder

1. Repetition-aware bi-gram
2. GRU-based language model

Overview

Related works

Markov model (HMM・n-gram) [Paulus+, 09]

What? Statistical language model considering temporal dependencies

Point: Simple architecture and good performance

Deep language model (RNN) [Sigtia+, 15]

What? Evaluate musical naturalness with recurrent neural networks

Point: High expressive power and easy implementation

Problem☹: It's hard to learn tatum-level musical structure due to the frame-level modeling

Related works

Cold fusion [Sriram+, 18]

What? Logit output of a pretrained language model is used in the training phase

Point Easy implementation and fastness at run time

Bayesian inerence [Ueda, 19]

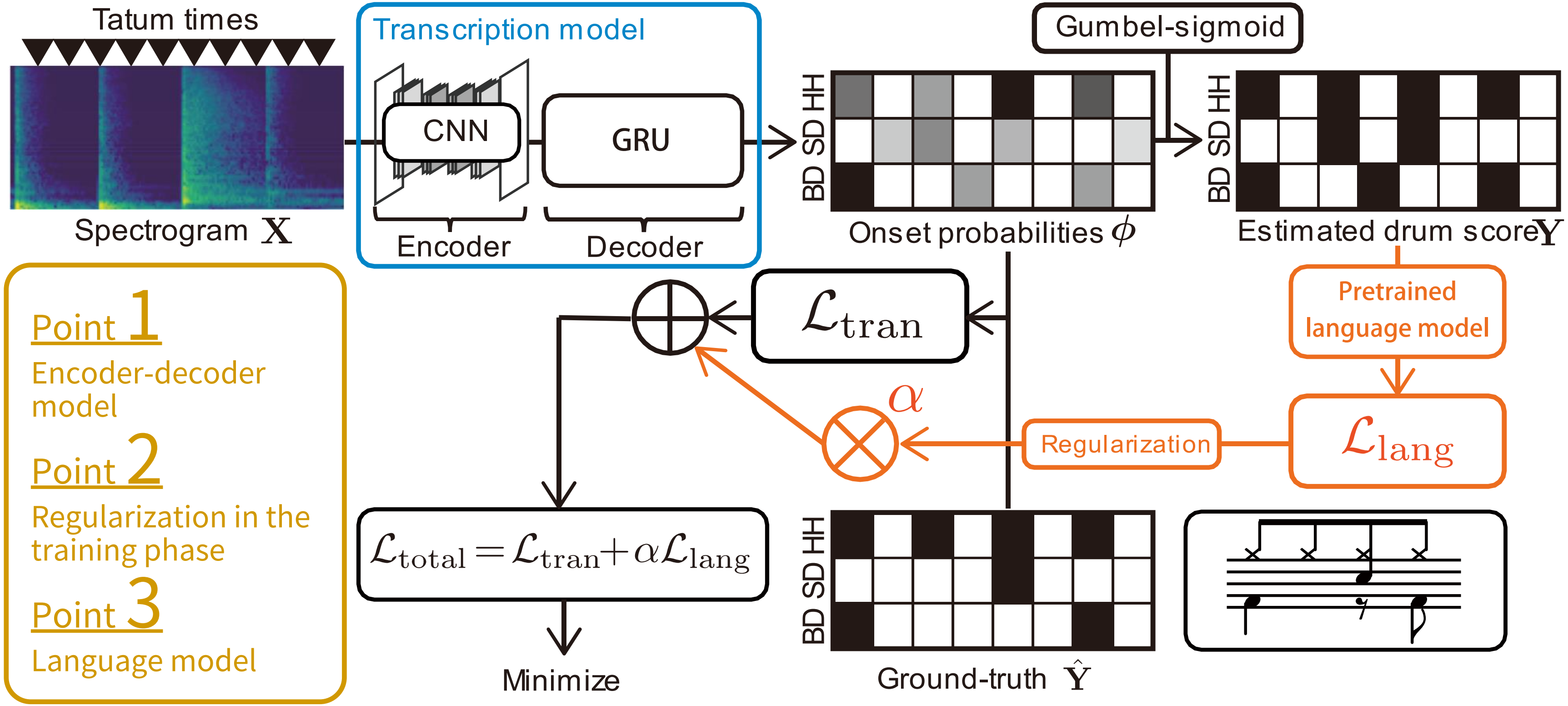
What? A VAE-based language model is used as a prior of the NMF-based transcription model

Point Flexible prior distribution

Point: **There are few studies to integrate a DNN-based language model into a DNN-based transcription model in ADT**

Overview

Proposed method



Point 1

Encoder-decoder model

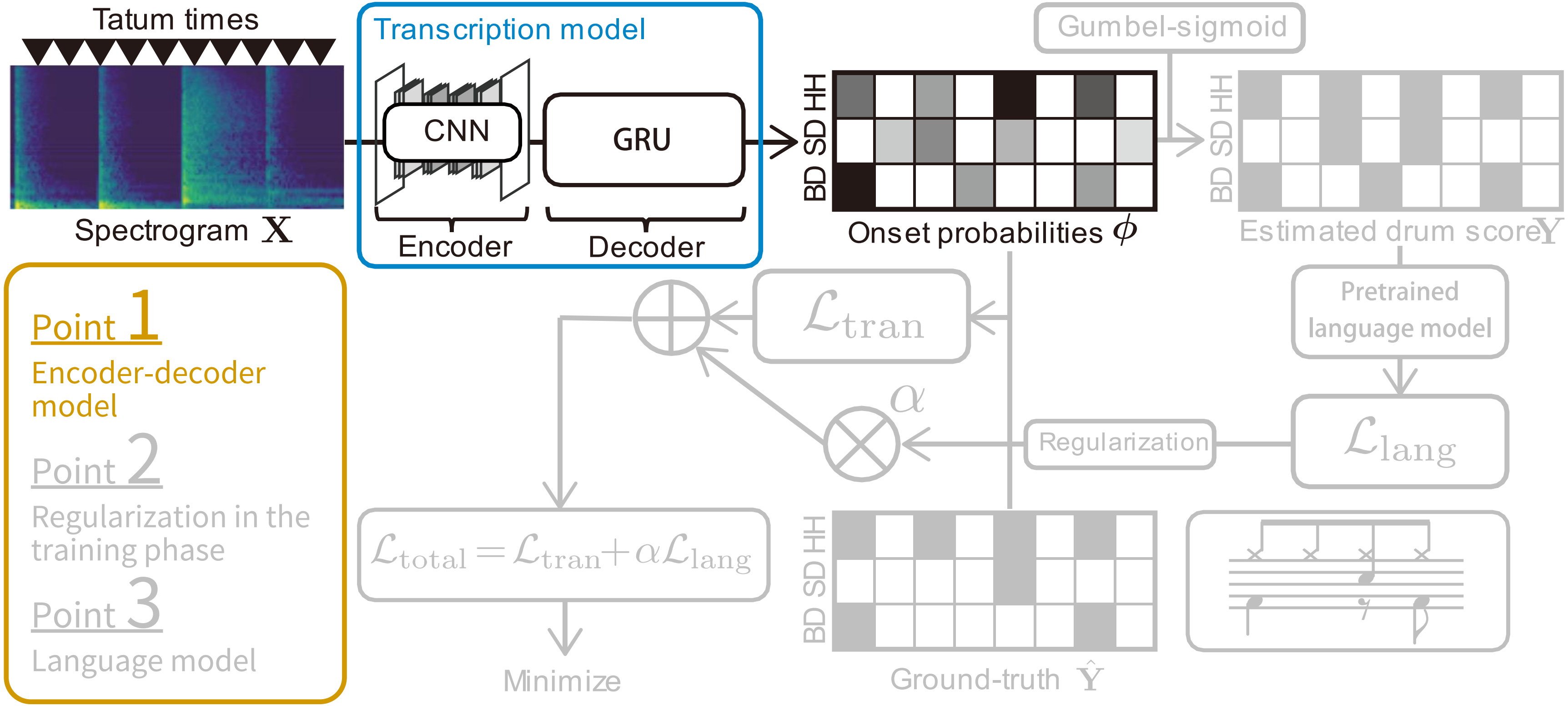
Point 2

Regularization in the training phase

Point 3

Language model

Proposed method



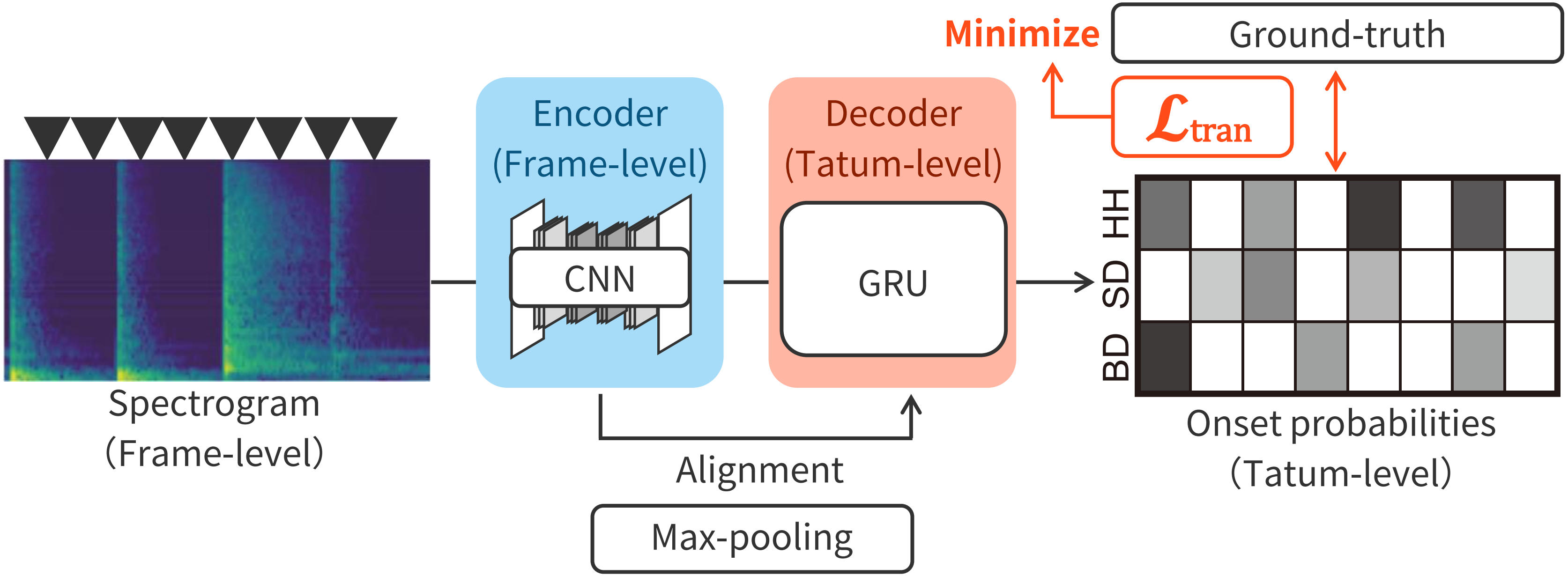
Point 1
Encoder-decoder model

Point 2
Regularization in the training phase

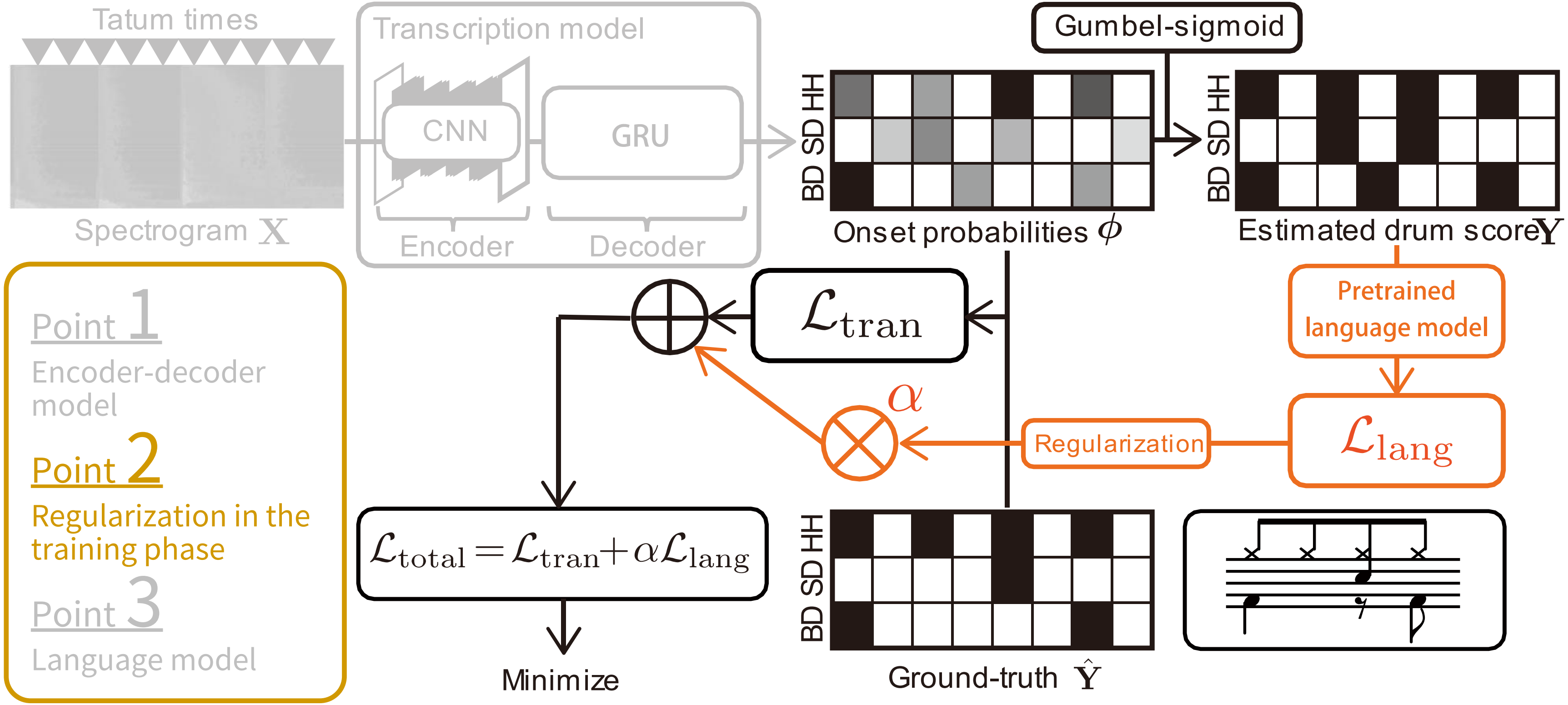
Point 3
Language model

Proposed method

Point 1: Encoder-decoder model



Proposed method



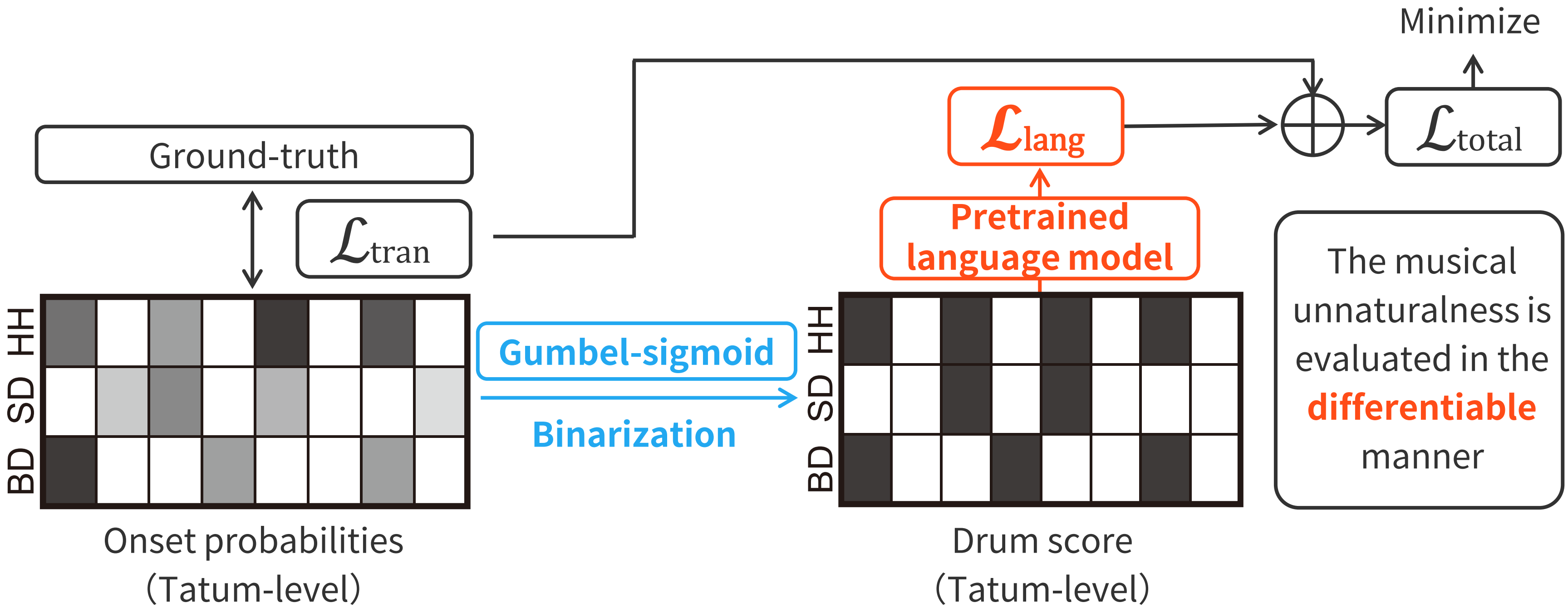
Point 1
Encoder-decoder model

Point 2
Regularization in the training phase

Point 3
Language model

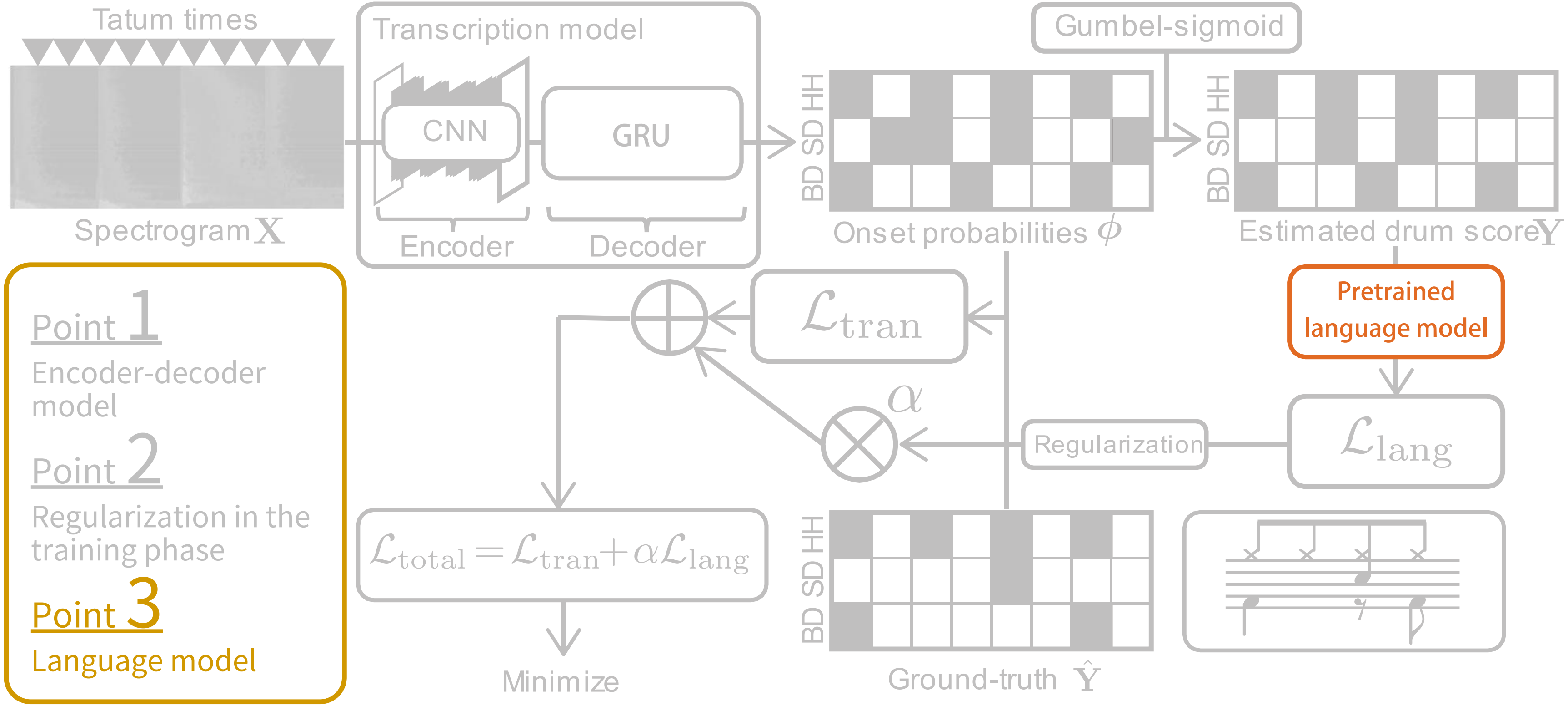
Proposed method

Point 2: Regularization in the training phase



The musical unnaturalness is evaluated in the **differentiable** manner

Proposed method



Point 1
Encoder-decoder model

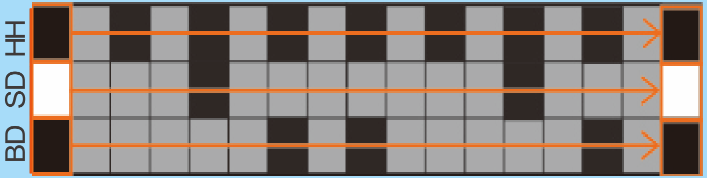
Point 2
Regularization in the training phase

Point 3
Language model

Proposed method

Point 3: Design of the language models

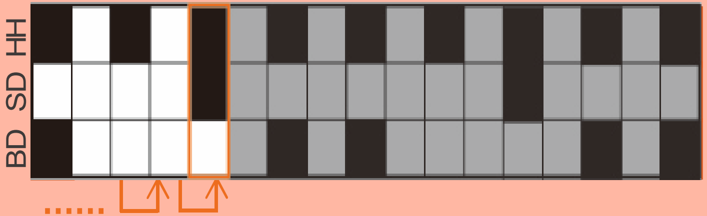
Repetition-aware bi-gram



Skip type (16 tatums) transitions

$$p(Y_{:,m}|Y_{:,1:m-1}) = \prod_{k=1}^{K} p(Y_{k,m}|Y_{k,m-16})$$

GRU



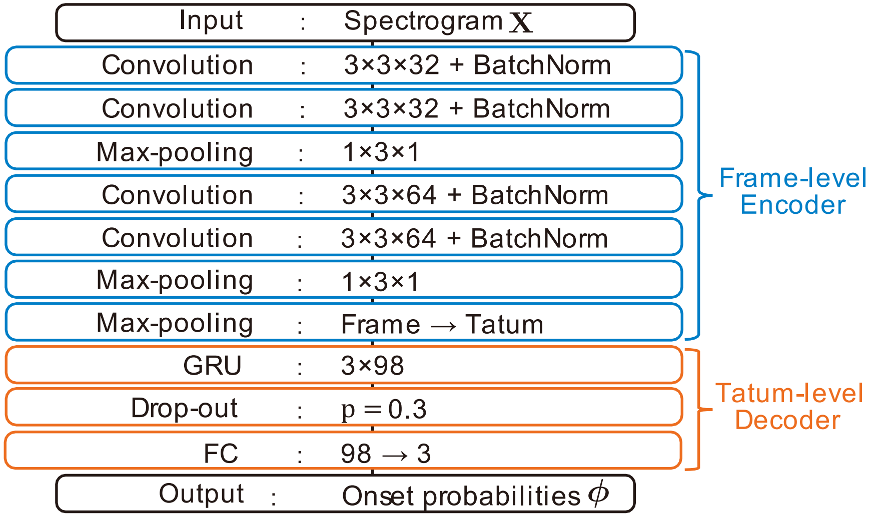
Transitions from the first tatum

$$p(Y_{:,m}|Y_{:,1:m-1})$$

Overview

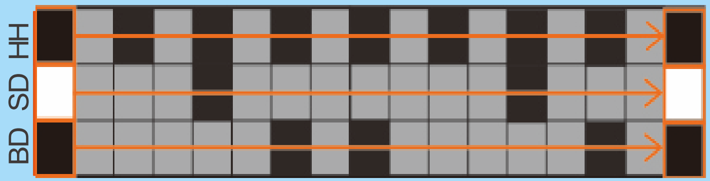
Experiments

Dataset (Transcription)	RWC popular music database (65 songs)
Dataset (Language model)	Jpop & Beatles (512 songs)
Test / Validation	10-fold cross validation (test) 15% of training data (valid)
Data augmentation	Spleeter
Architecture	CRNN as in the right figure
Audio features	Mel spectrogram (80bands)
Hyperparameters	Optuna
Measurement	Precision / Recall / F-measure
Beat estimation	Madmom ($\mathcal{F} = 96.4\%$)



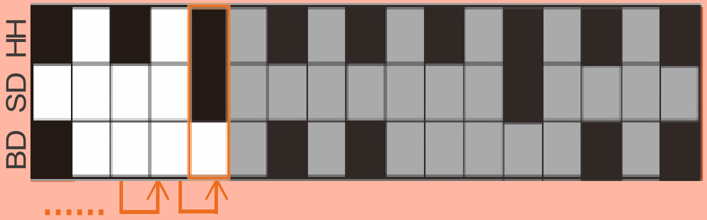
Language model	Perplexity
Bi-gram	1.51
GRU	1.44

Repetition-aware bi-gram



Skip type (16 tatum) transitions

GRU



Transitions from the first tatum

GRU was better than bi-gram

Experiments

Weighting factor of the language model

		Madmom			Ground-truth		
		$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
State-of-the-art →	CRNN [7]	70.8	77.4	65.9	71.0	77.6	66.1
Without regularization →	CRNN	78.9	86.3	73.1	79.3	86.7	73.3
With regularization →	+ Bi-gram ($\alpha = 0.068$)	81.4	84.7	79.1	80.8	83.7	78.8
With regularization →	+ GRU ($\alpha = 0.055$)	**81.6**	84.0	80.2	**81.1**	83.2	79.7

This experiment showed that...

1. The frame-to-tatum outperformed the SoTA method by **8 points**
2. The language model-based regularization outperformed the non-regularized method by **2 points**
3. The **GRU-based regularization** had much improvement than the bi-gram-based regularization

[7] Vogl, Richard, et al. "Drum Transcription via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks." *ISMIR*. 2017.

Experiments

RWC-MDB-P-2001 No.88 ($\mathcal{F} = 58.3\% \rightarrow 79.9\ \%$)

Summary

Points

Viewpoint　Frame-to-frame methods predict musically-unnatural drum patterns

Keypoint　Tatum-level language model-based regularized training

Experiments

	:	The frame-to-tatum architecture improved **about 8 points**
F-measure	:	The regularization improved **about 2 points**
	:	**GRU** has much improvement than bi-gram

Future works

Work1	:	Dealing with **Fill-ins**
Work2	:	Learn other than three main parts such as **symbals** and **toms**
Work3	:	Capturing **global structure** with self-attention mechanism