ISMIR

Sequence-to-Sequence Network Training Methods for Automatic Guitar Transcription with Tokenized Outputs

NAGOYA UNIVERSITY

Sehun Kim, Kazuya Takeda, Tomoki Toda
Nagoya University, Japan

✓ Introduction

□ Automatic music transcription

- Task to automatically generate musical symbol from audio
- **Objective**: generate playable sheet music

□ Tokenization of music score

- A way to represent a musical score as a series of note events
- Widely used for tasks such as AMT and music generation

□ Recent approach : Seq-to-seq network

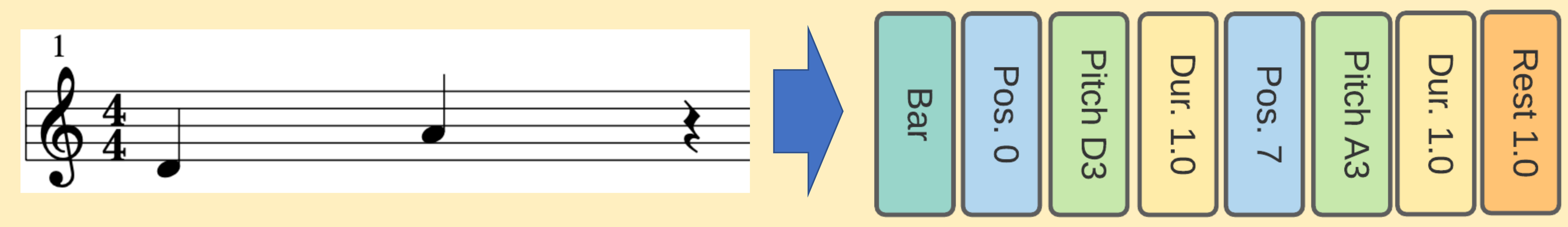
☺ **Learning a musical language model** to achieve **musical context-aware** automatic music transcription

☹ Performance tends to be **extremely poor when the amount of training data is small**

✓ Related works

□ Tokenization : REMI [Hwang+ 2020]

- Express the location of a note in position. First introduced in automatic music generation task



- It requires **large amount of data** to properly train
  - **Guitar has less available data** than piano.

□ AMT system based on Transformer

- A system that only predicts token sequence [Hawthorne+ 2021]
- A system that predicts both token sequence and frame-level pianoroll [Chen+ 2022]

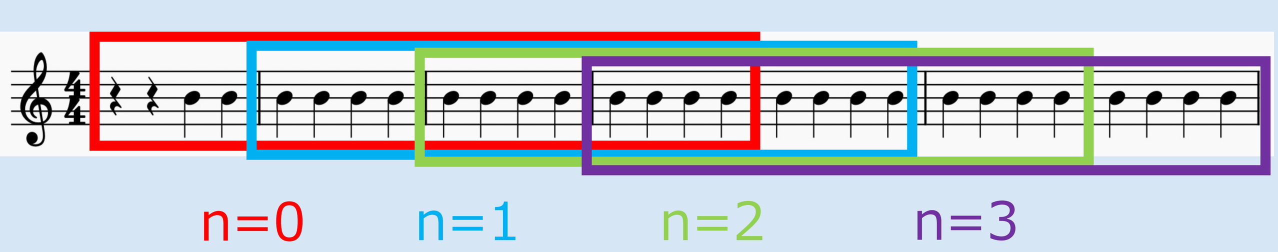
✓ Contribution of this research

1. Proposal of **two data augmentation methods** to increase the amount of training data
2. Proposal of **Hybrid CTC-Attention model** for automatic guitar transcription which improves transcription performance especially when training with small amounts of data

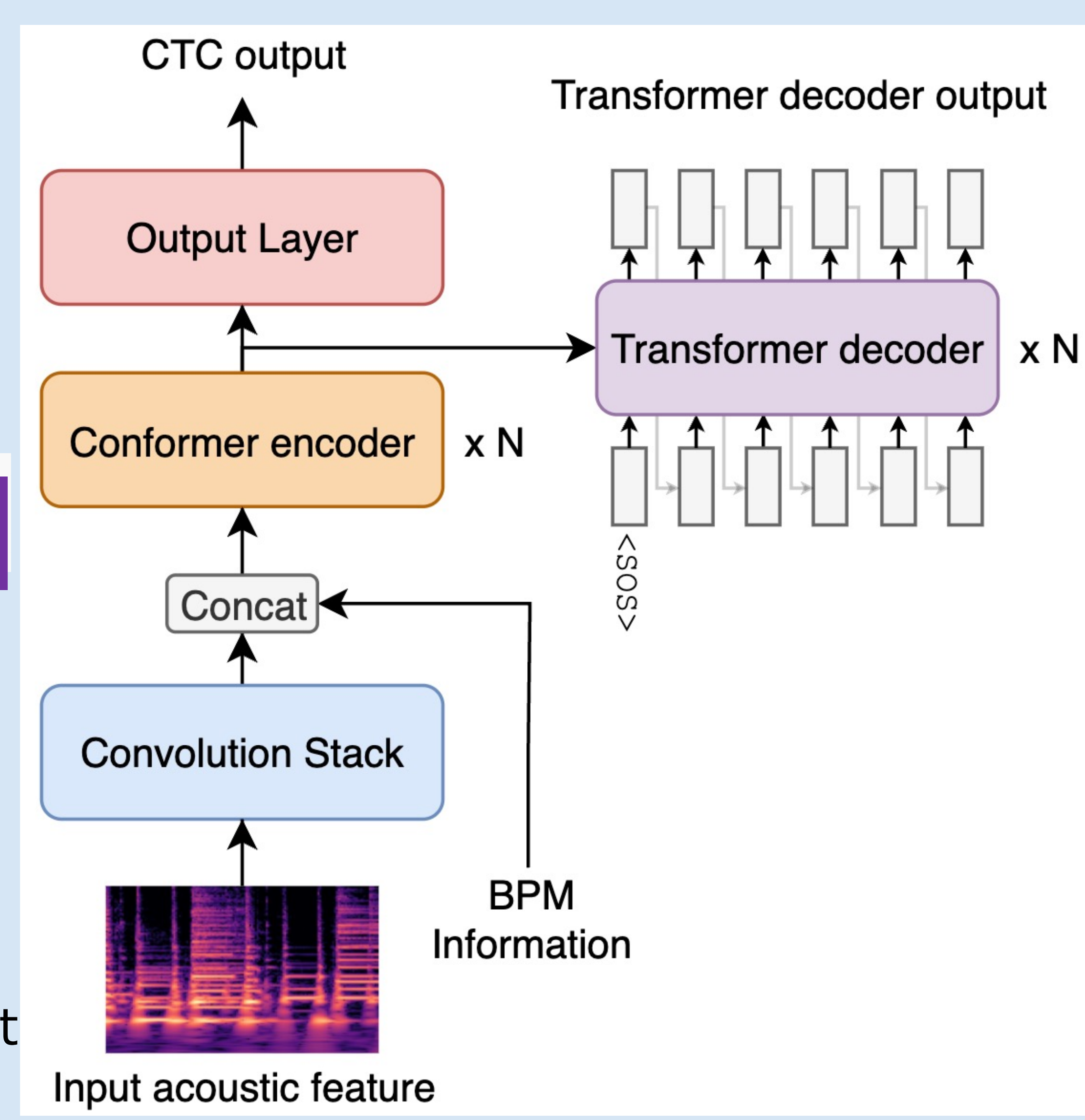
✓ Proposed method

□ Data augmentation

- **Bar overlap (BO)**
  - Preserves musical structure by taking segments in units of bars instead of fixed length, and **shift the window**



- **Pretraining with Synthetic audio-MIDI pair (PT)**
  - Using an **oscillator from MIDI-only data** to create a large amount of synthetic audio-MIDI pair data
  - **Pretrain** using an artificially created dataset and finetune using a real guitar dataset



□ Hybrid CTC-Attention model

[Watanabe+, 2017]

- Basic structure is similar to the Conformer-Transformer speech recognition model
- **Multi-task learning** with two types of token estimation with Transformer decoder output and CTC output from Conformer encoder

$$\mathcal{L}_{total} = \alpha\underbrace{\mathcal{L}_{\mathrm{CTC}}}_{\text{CTC loss}} + \underbrace{\mathcal{L}_{\mathrm{Transformer}}}_{\text{Cross entropy loss}}$$

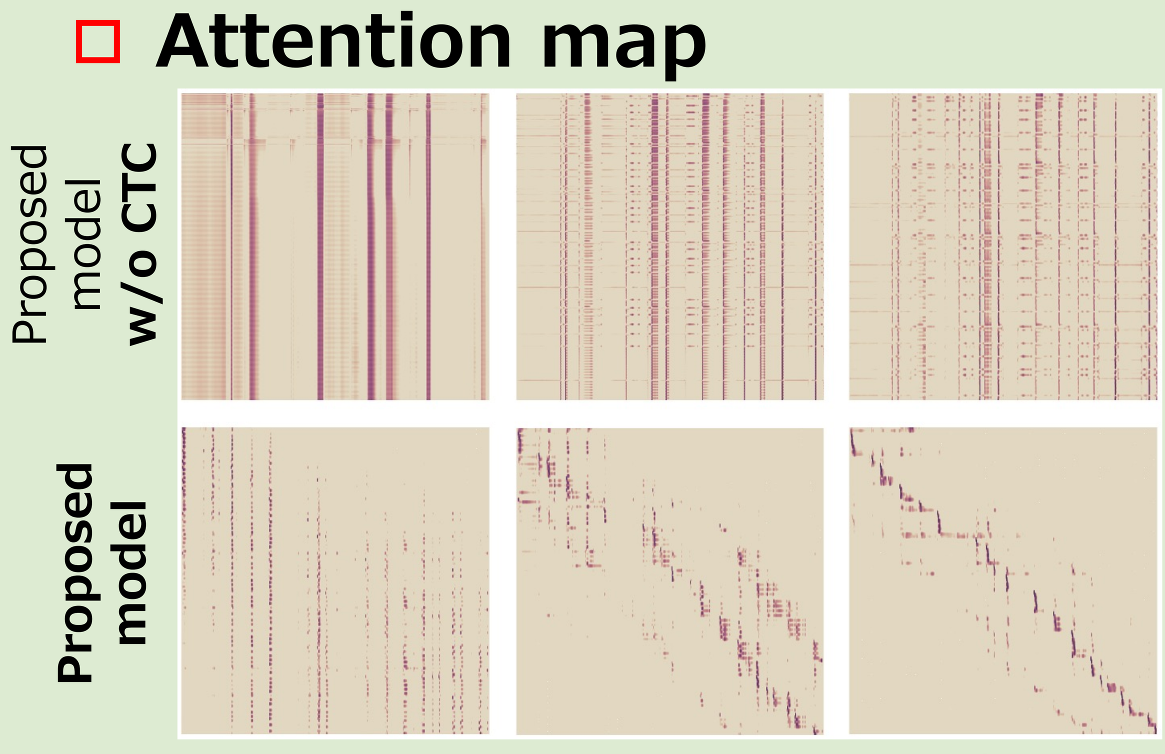
→ The **monotonic alignment constraint** of the CTC helps the attention mechanism to learn the proper alignment between input and output, especially when there is only a small amount of data

✓ Experimental evaluation

□ Dataset

- Data used for data augmentation : Classic guitar MIDI archive
  - **MIDI-only** classical guitar data set
  - More than **20 hours** of data in total
- A dataset with **real guitar** recordings : GuitarSet [Xi+, 2018]
  - An acoustic guitar dataset composed of Audio-MIDI pairs
  - Six performers, about **3 hours** of data in total

□ Attention map



- Experiments done using **GuitarSet only** to confirm the effectiveness of CTC when training with only a **small amount of data**
- We confirmed that the introduction of **CTC helps Attention mechanism to learn proper alignment**

□ Effect of data augmentation

	Encoder output		Decoder output	
Method	F1 ↑	TER ↓	F1 ↑	TER ↓
No data augmentation	0.363	0.469	0.526	0.712
Proposed (BO)	0.512	0.365	0.699	0.441
Proposed (PT)	0.555	0.388	0.630	0.497
Proposed (BO+PT)	**0.666**	**0.307**	**0.803**	**0.335**

□ Effect of Hybrid CTC-Attention model

	Encoder output		Decoder output	
Method	F1 ↑	TER ↓	F1 ↑	TER ↓
Baseline [Chen+ 2022]	0.767	-	0.603	0.589
Proposed w/o CTC	-	-	0.784	0.345
Proposed	0.666	0.307	**0.803**	**0.335**