**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
  - 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

**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**

<table>
<thead>
<tr>
<th>Method</th>
<th>Encoder output</th>
<th>Decoder output</th>
</tr>
</thead>
<tbody>
<tr>
<td>No data augmentation</td>
<td>F1 ↑ 0.363 0.469</td>
<td>TER ↓ 0.526 0.712</td>
</tr>
<tr>
<td>Proposed (BO)</td>
<td>0.512 0.365</td>
<td>0.699 0.441</td>
</tr>
<tr>
<td>Proposed (PT)</td>
<td>0.555 0.388</td>
<td>0.630 0.497</td>
</tr>
<tr>
<td>Proposed (BO+PT)</td>
<td>0.666 0.307</td>
<td>0.803 0.335</td>
</tr>
</tbody>
</table>

**Effect of Hybrid CTC-Attention model**

<table>
<thead>
<tr>
<th>Method</th>
<th>Encoder output</th>
<th>Decoder output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [Chen+ 2022]</td>
<td>F1 ↑ 0.767</td>
<td>TER ↓ 0.603 0.589</td>
</tr>
<tr>
<td>Proposed w/o CTC</td>
<td>- -</td>
<td>0.784 0.345</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.666 0.307</td>
<td>0.803 0.335</td>
</tr>
</tbody>
</table>