High-Resolution Violin Transcription using Weak Labels

We present the Multi-Stream Conformer (MUSC), a SOTA violin transcriber that converts 44.1 kHz raw audio into MIDI with 5.8ms time- and 10-cent frequency-resolution, and without requiring frame-wise labels during training!

Nazif C. Tamer, Yigitcan Özer, Meinard Müller, Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain, International Audio Laboratories Erlangen, Germany

CNNs and Conformer blocks convert raw audio into 4 representations, then postprocessing creates the MIDI with pitch bends.

MIDI transcription with pitch bends

Postprocessing

Constrained Viterbi Pitch Estimation

Notes

Onsets (semitone)

Offsets (semitone)

Frames (semitone)

10s (10 cents)

Multi-Stream Conformer

Conformer Blocks

Conformer Blocks

Conformer Blocks

Conformer Blocks

x4

Conformer Blocks

Stack

Dilated CNN

FC

Conformer Blocks

Conformer Blocks

Conformer Blocks

Conformer Blocks

x16

5.8 ms rate violin representations

44.1 kHz waveform

Training

As a large-scale violin dataset with frame-wise labels do not exist, MUSC generates its frame-wise training labels by aligning its own onset, offset, and frame feature representations with music score.

Constrained Viterbi Pitch Estimation

Audio-Score Alignment

Notes

Onsets

Offsets

Frames

10s

Multi-Stream Conformer

Conformer Blocks

Conformer Blocks

Conformer Blocks

Conformer Blocks

x4

Conformer Blocks

Stack

Dilated CNN

FC

Conformer Blocks

Conformer Blocks

Conformer Blocks

Conformer Blocks

x16

5.8 ms rate violin representations

44.1 kHz waveform

Dataset: It is trained on 120 violin etudes from three books & their unaligned YouTube recordings.

<table>
<thead>
<tr>
<th>Method</th>
<th>Etudes</th>
<th>Players</th>
<th>Performances</th>
<th>Duration (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paganini, Op. 1</td>
<td>24</td>
<td>10</td>
<td>235</td>
<td>13:00</td>
</tr>
<tr>
<td>Wohlfahrt, Op. 45</td>
<td>60</td>
<td>6</td>
<td>506</td>
<td>11:36</td>
</tr>
<tr>
<td>Kayser, Op. 20</td>
<td>41</td>
<td>8</td>
<td>280</td>
<td>09:48</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>22</td>
<td>1021</td>
<td>34:23</td>
</tr>
</tbody>
</table>

The released dataset can be found on GitHub, with the frame-wise alignments generated by our model.

Transcription

Compared with MT3 and Basic Pitch on two datasets. (URMP is involved in MT3 training set.)

<table>
<thead>
<tr>
<th>URMP</th>
<th>Bach10</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>MUSC</td>
<td>86.5</td>
</tr>
<tr>
<td>MT3</td>
<td>79.1</td>
</tr>
<tr>
<td>BP</td>
<td>58.8</td>
</tr>
</tbody>
</table>


Tests yield SOTA performance for two proxy tasks: Violin Transcription & Pitch Estimation

Transcription

Compared with MT3 and Basic Pitch on two datasets. (URMP is involved in MT3 training set.)

<table>
<thead>
<tr>
<th>URMP</th>
<th>Bach10</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>MUSC</td>
<td>98.3</td>
</tr>
<tr>
<td>vMUSC</td>
<td>98.6</td>
</tr>
<tr>
<td>CREPE</td>
<td>96.4</td>
</tr>
<tr>
<td>vCREPE</td>
<td>97.3</td>
</tr>
<tr>
<td>YIN</td>
<td>95.3</td>
</tr>
<tr>
<td>pYIN</td>
<td>97.2</td>
</tr>
<tr>
<td>SWIPE</td>
<td>97.2</td>
</tr>
</tbody>
</table>