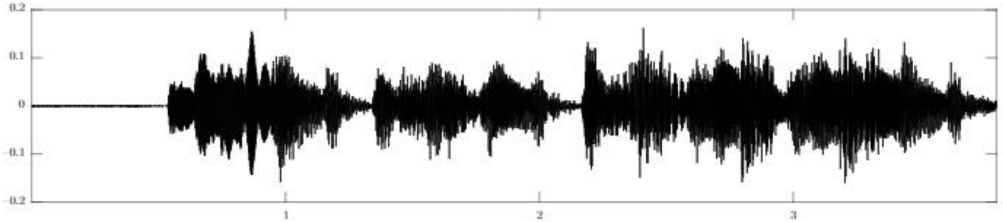
Automatic Music Audio-to-Score Transcription with Deep Neural Networks

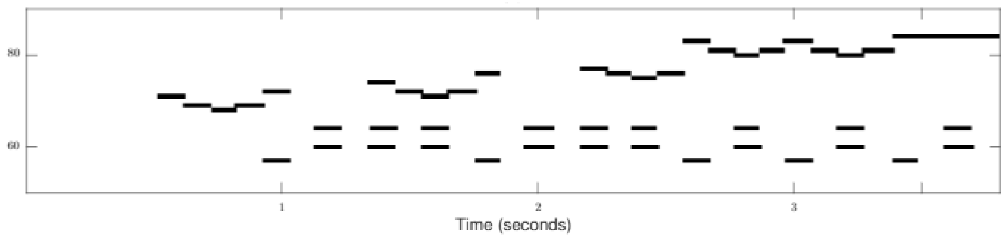
Lele Liu
Supervisors: Emmanouil Benetos, Veronica Morfi, and Simon Dixon

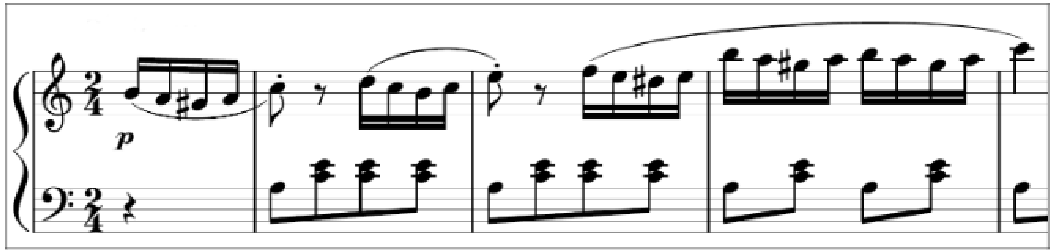
AIM forum, Oct 29, 2021

AIM AI + MUSIC
China Scholarship Council
UK Research and Innovation
centre for digital music
Queen Mary University of London

Problem Definition







Multi-pitch detection

Output format: piano-roll

Score transcription

Output format: music score

Time-Frequency Representation Comparison

- A comparison between **different time-frequency representations** (STFT, Mel spectrogram, CQT, HCQT and VQT[1]) and **their different parameters**

- Performance evaluated on a **multi-pitch detection task** with a **Convolutional Recurrent Neural Network** (CRNN) network architecture

- Results evaluated on MIREX multi-pitch detection metric on the synthesized dataset

- **VQT** shows the best performance, **with a γ value of 20 and 8 octaves × 60 bins per octave** in the frequency axis

Table 2. F-measure of piano-roll prediction on different input representations and models. F_f: frame-level, F_{on}: note-level onset only, F_{onoff}: note-level onset and offset.

Input representations/Models	F_f	F_{on}	F_{onoff}
STFT	89.5	81.0	61.7
Mel Spectrogram	89.0	82.1	63.0
CQT	**91.9**	85.4	67.4
HCQT	91.0	84.1	65.3
VQT	**91.9**	**85.7**	**68.5**

[1] STFT: Short-Time Fourier Transform; CQT: Constant-Q Transform; HCQT: Harmonic Constant-Q Transform; VQT: Variable-Q Transform

Proposed Score Representation

- **LilyPond representation** vs. Proposed **Reshaped representation**
- Model combined with a convolutional network and two attentional sequence-to-sequence models for right and left hand score prediction.

- Model performance tested on the synthesized dataset.
- The Reshaped representation also outperforms the LilyPond representation in terms of the time and memory resources required (around **7 times faster** and **half** the memory)

Music score:

LilyPond representation:

g'4<e'~a'>8<e'g'>8<c'e'g'>4<c'f'a'>4

Reshaped representation:

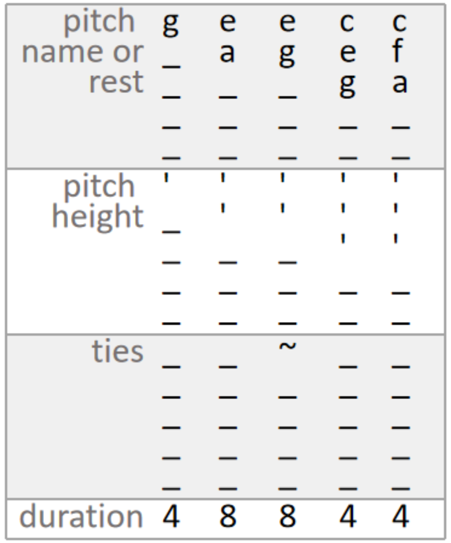


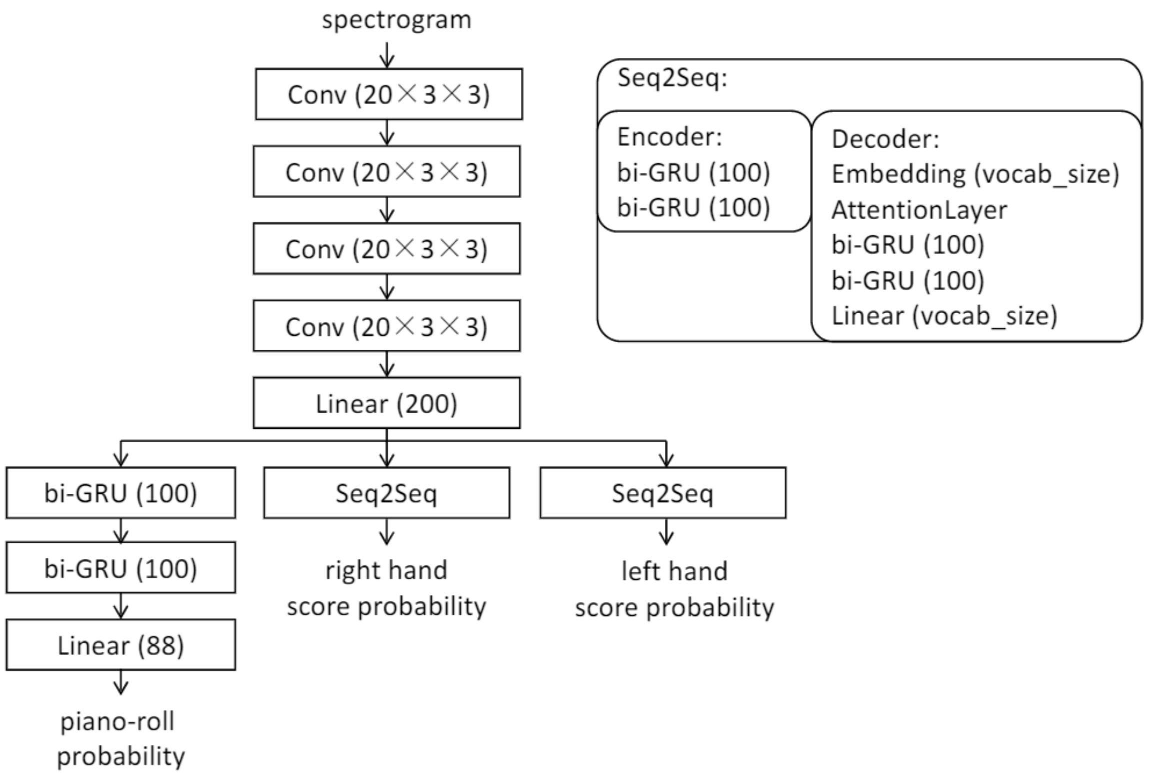
Table 3. Word error rates and MV2H results in percentage for different score representations.

WER	wer_{right}		wer_{left}		wer
LilyPond	38.0		39.0		38.5
Reshaped	**37.8**		**34.5**		**36.2**
MV2H	F_p	F_{voi}	F_{met}	F_{val}	F_{MV2H}
LilyPond	66.7	**90.3**	94.8	93.2	86.3
Reshaped	**69.6**	89.7	94.8	**93.7**	**86.9**

MV2H: Andrew Mcleod and Mark Steedman. Evaluating Automatic Polyphonic Music Transcription. In ISMIR, pages 42–49, 2018.

Joint Multi-pitch Detection & Score Transcription

- Model architecture (attention implementation follows Bahdanau et al. 2015):



- The Joint model generally outperforms the single task models.

Table 4. Performances on single-task and multi-task models.

F-measure	F_f		F_{on}		F_{onoff}
Piano-roll Only	86.4		**67.6**		52.0
Joint	**88.0**		66.7		**53.6**
WER	wer_{right}		wer_{left}		wer
Score-only	37.8		**34.5**		**36.2**
Joint	**37.6**		35.3		36.5
MV2H	F_p	F_{voi}	F_{met}	F_{val}	F_{MV2H}
Score-only	69.6	89.7	94.8	93.7	86.9
Joint	**71.1**	**90.8**	**94.9**	**94.4**	**87.8**

ACPAS Dataset

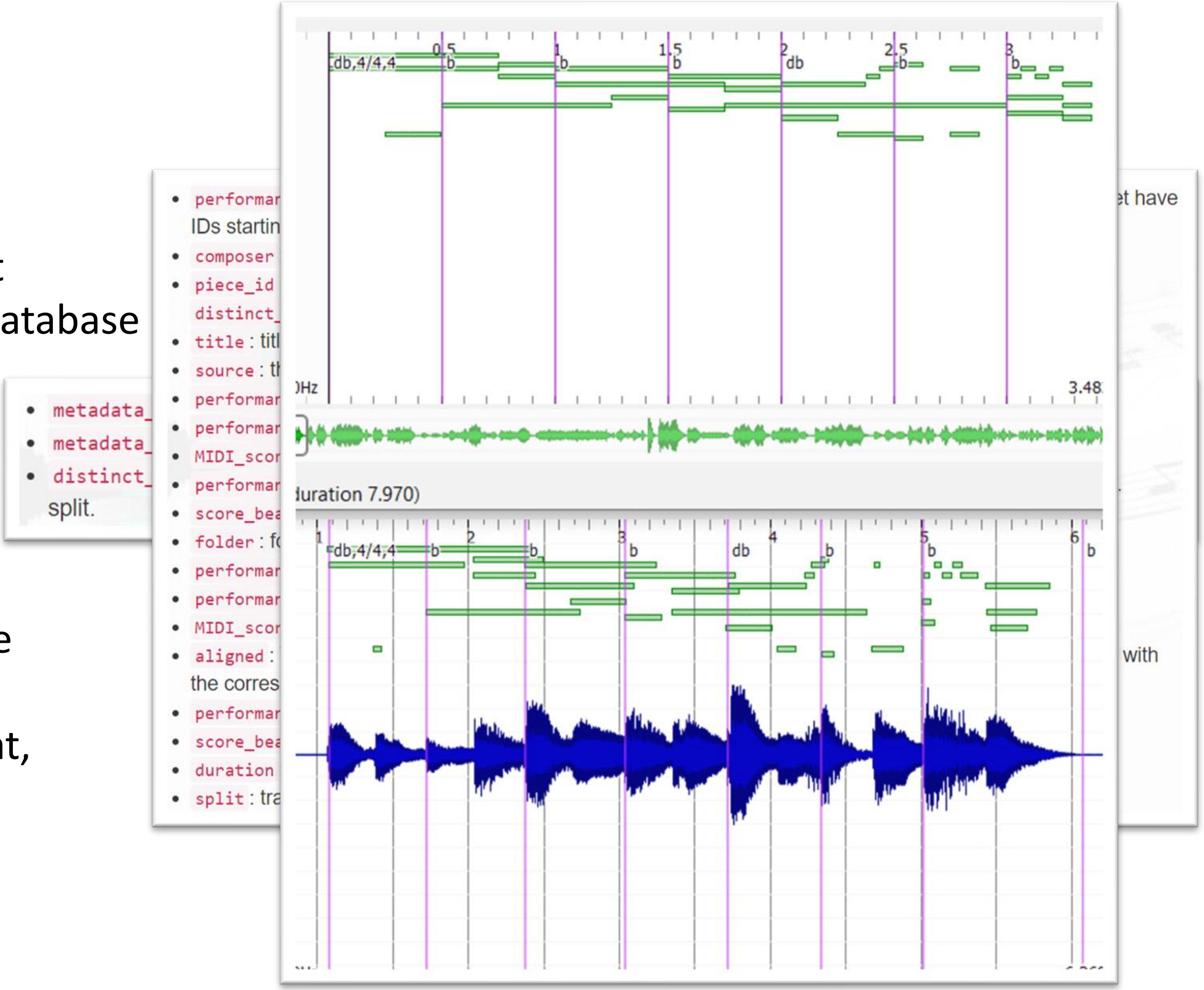
- 497 distinct music scores aligned with 2189 audio performances.
- Currently the largest dataset for audio-to-score transcription, to our knowledge.
- Aligned performance audio, performance MIDI and MIDI scores, together with beat, key signature, and time signature annotations
- A train/validation/test split with **no piece overlap** and in line with splits in other automatic music transcription datasets

Statistics across subsets and splits

Subset/Split	Distinct Pieces	Performances	Duration (hours)
Real recording	215	578	49.0
Synthetic	497	1611	130.8
train	359	1523	127.7
validation	49	184	11.2
test	89	482	40.9
Total	497	2189	179.8

Dataset Content

- ❑ We collected data from three sources
  - the MAPS[1] and A-MAPS[2] dataset
  - the Classical Piano MIDI[3] (CPM) database
  - the ASAP[4] dataset
- ❑ Two subsets
  - Real recording subset
  - Synthetic subset
- ❑ Synthesis process
  - Four different piano fonts in Native Instrument Kontakt Player[5]
  - Monaural audio files in .wav format, 44.1kHz, 16 bit



Next steps

- ❑ Automatic performance MIDI to quantized MIDI conversion (internship project)
- ❑ Modelling longer sequence using enhanced model architecture and larger dataset
- ❑ Cross-domain audio-to-score transcription

Feedback and contact:

Email: lele.liu@qmul.ac.uk
Website: https://cheriell.github.io

Thank you for your attention!