

2021 Intelligent Sensing Winter School
- AI for sound perception

CIS centre for
intelligent sensing



Joint multi-pitch detection and score transcription for piano music recordings

Lele Liu

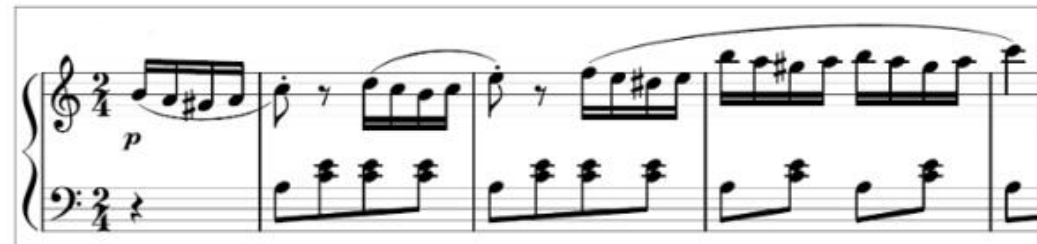
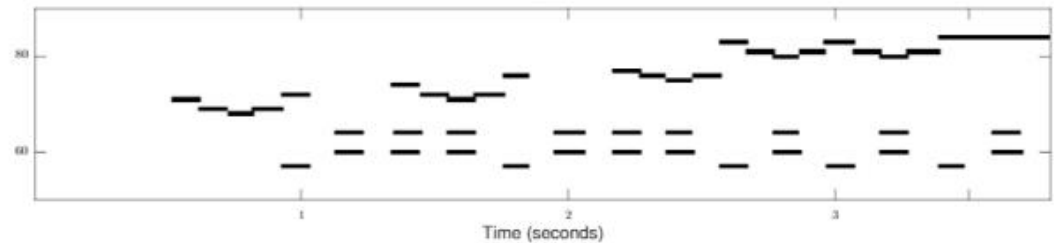
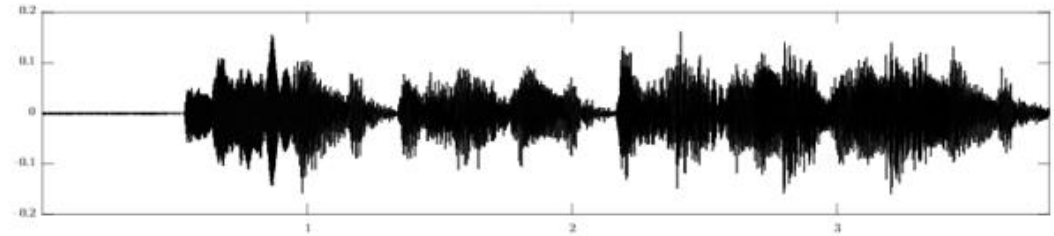
Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London

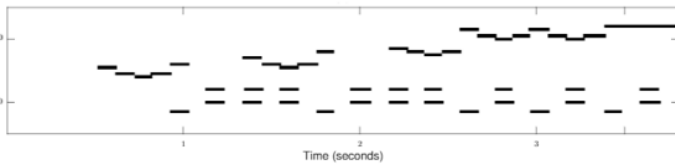
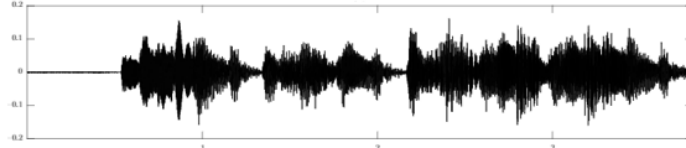


centre for digital music

Automatic Music Transcription

- ▶ **Automatic music transcription (AMT)** is the task of transcribing a human- or machine-readable musical score from a music recording using computer algorithms.
- ▶ It is common to get a **piano-roll format** transcription (multi-pitch detection), or a **score format** transcription (score transcription)





Multi-pitch detection

Output format: piano-roll

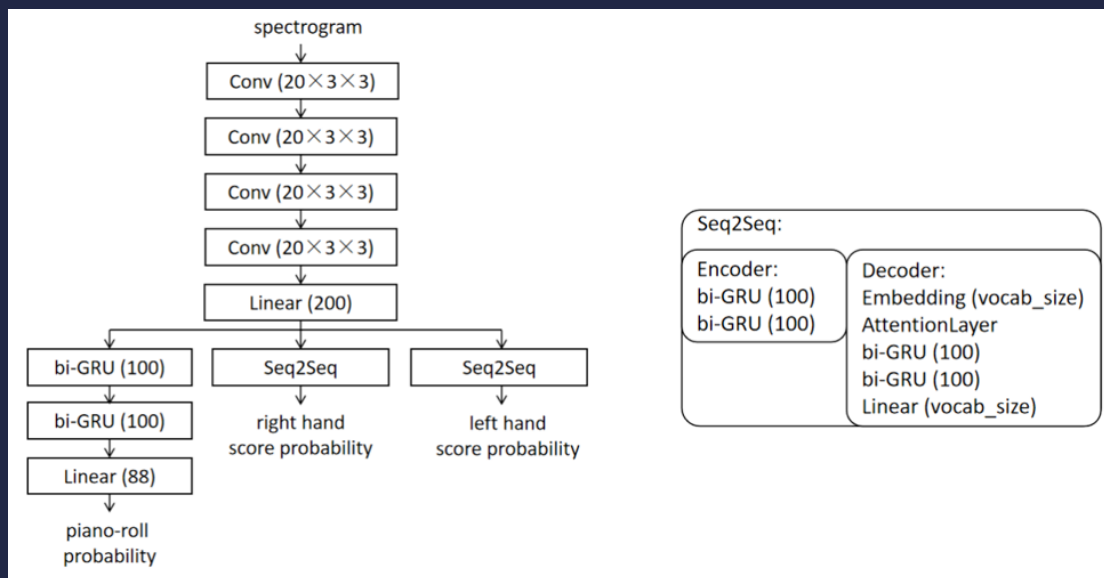


Score transcription

Output format: music score

Joint multi-pitch detection and score transcription

Multitask model



- ▶ Four Convolutional layers for feature extraction
- ▶ Bi-RNN layers on top of the CNN stack for piano-roll prediction
- ▶ Sequence-to-sequence models in parallel with the RNN layers for score transcription prediction. For piano music, we separate notes for right and left hand as in two sequences.

Experimental data

- ▶ Synthesized dataset with scores collected from the MuseScore website
- ▶ Audio files synthesized using four piano models using the Native instrument Kontakt Player
- ▶ Three piano models for train/validation and all four piano models for testing
- ▶ Train:valid:test = 8:1:1

Table 1. Dataset Statistics. For polyphony levels, the numbers out of brackets are calculated without adding piano pedals, and the numbers in brackets are calculated with piano pedals.

Number of music pieces	210
Total hours	9.62×4 piano models
Total notes	222,219
Use of piano pedal	29% (61 pieces)
Maximum polyphony level	13 (26)
Average polyphony level	2.87 (3.21)
Time signatures	4/4, 3/4, 5/4, 6/8, 9/8, etc.
Key signatures	all 12 key signatures

Experiments on different input spectrograms

- ▶ A comparison on different types and parameters of audio spectrograms for model input:
 - ▶ Short-Time Fourier Transform (STFT)
 - ▶ Mel Spectrogram
 - ▶ Constant-Q Transform (CQT)
 - ▶ Harmonic Constant-Q Transform (HCQT)
 - ▶ Variable-Q Transform (VQT)
- ▶ The latter three spectrograms shows better performance
- ▶ Using more frequency bins in spectrograms tend to achieve better performance
- ▶ Multi-task model outperforms single-task model


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. The last two models use VQT as input, and are evaluated on all four pianos in the dataset.

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
Piano-roll only	86.4	67.6	52.0
Joint	88.0	66.7	53.6

Experiments on output score representation

- ▶ We compare between the LilyPond representation and a Reshaped representation
- ▶ Reshaped representation outperforms the original LilyPond representation in transcription accuracy as well as in terms of the time and memory resources required (around 7 times faster and half the memory)
- ▶ Joint model achieves higher accuracy than single-task score transcription.

Music score:



LilyPond representation:

`g'4<e'^a>8<e'g'>8<c'e'g'>4<c'f'a'>4`

Reshaped representation:

pitch	g	e	e	c	c
name or	-	a	g	e	f
rest	-	-	-	g	a
pitch	'	'	'	'	'
height	4	4	4	4	4
ties	-	-	~	-	-
duration	4	8	8	4	4

Table 3. Word error rates and MV2H results in percentage for different models. LilyPond: Score-only model with LilyPond representation; Reshaped: Score-only model with Reshaped representation; Joint: Joint model with Reshaped representation. Models evaluated on four pianos in the dataset.

WER	w_{right}	w_{left}	w_{er}		
LilyPond	38.0	39.0	38.5		
Reshaped	37.8	34.5	36.2		
Joint	37.6	35.3	36.5		
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
Joint	71.1	90.8	94.9	94.4	87.8

Transcription examples

Two musical staves showing piano and transcription. The top staff is a piano score with a tempo of quarter note = 70, marked *p* and *rit.* The bottom staff is a transcription of the same piece, showing the underlying harmonic structure with block chords and a simplified melodic line.

▶ Two transcription examples:

▶ Sample 1



▶ Transcription



▶ Sample 2



▶ Transcription



Two musical staves showing piano and transcription. The top staff is a piano score with a tempo of quarter note = 70, marked *p*. The bottom staff is a transcription of the same piece, showing the underlying harmonic structure with block chords and a simplified melodic line.

Thank you!

- ▶ This presentation is based on paper:
 - ▶ L. Liu, V. Morfi and E. Benetos, “Joint Multi-pitch Detection and Score Transcription for Polyphonic Piano Music,” IEEE International Conference on Acoustics, Speech and Signal Processing, Canada, Jun 2021.
- ▶ For any questions/suggestions, please feel free to contact:
- ▶ lele.liu@qmul.ac.uk