Towards an Integrated Approach for Expressive Piano Performance Synthesis from Music Scores

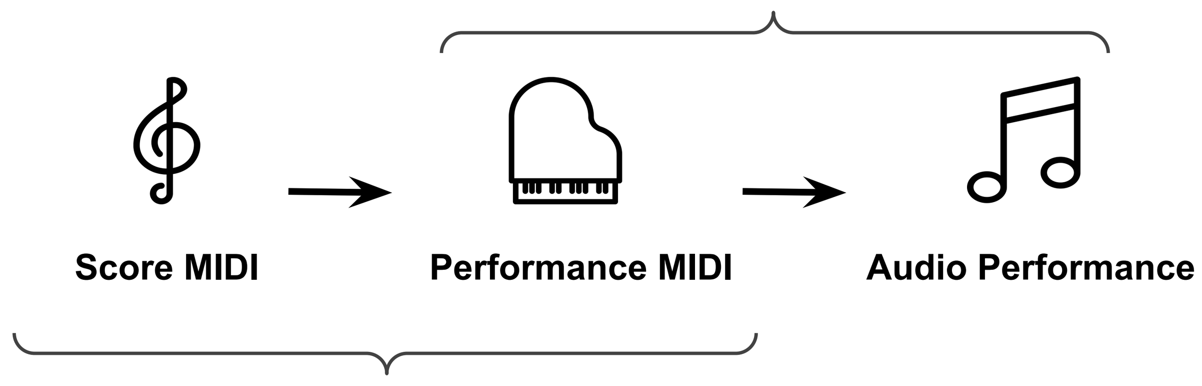
Jingjing Tang, Erica Cooper, Xin Wang, Junichi Yamagishi, Gyorgy Fazekas
Centre for Digital Music, Queen Mary University of London, UK & National Institute of Informatics, Japan

AI + MUSIC

Scan for Paper

Scan for Demo

Music Performance Synthesis



Objective: Generating expressive piano performances from **symbolic music representations (e.g., MIDI)**.

Challenges

- ❑ Difficulties in accurately aligning scores, performance MIDI, and recorded audio.
- ❑ Challenges in precisely adjusting dynamics, articulation, and timing variations.
- ❑ Limited adaptability to unseen compositions, instruments, and recording environments.

Related Works

Expressive Performance Rendering (EPR):

- ❑ RNNs, GNNs, GANs, Diffusion models.
- ❑ **Transformer**-based models (MIDI tokenisation)

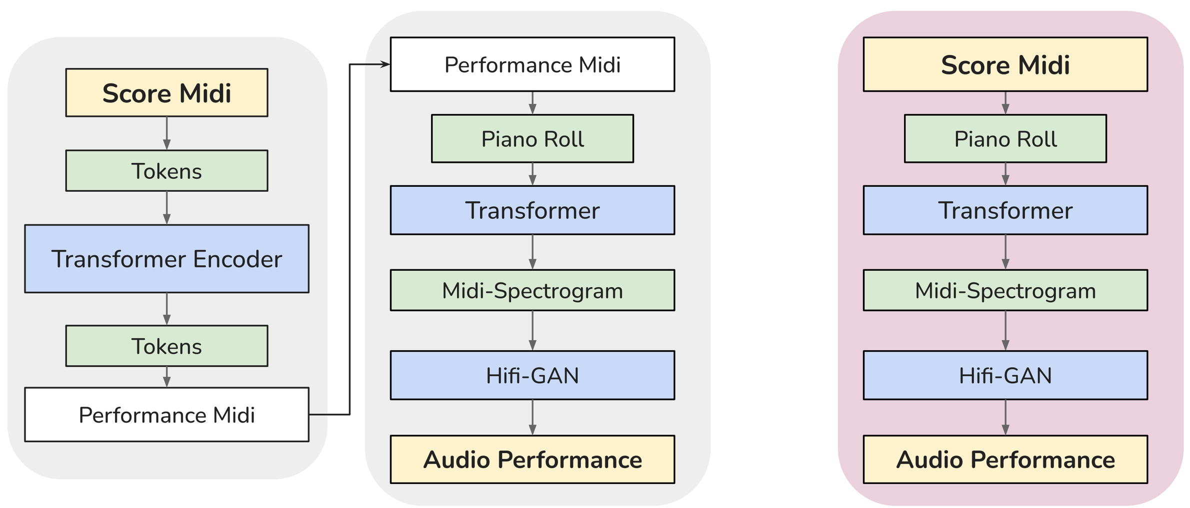
Expressive Performance Synthesis (EPS):

- ❑ Differentiable Digital Signal Processing (DDSP) models
- ❑ Adaptation from **Text-to-Speech** (TTS) models

Integrated System:

- ❑ MIDI-DDSP: Multi-instrument, monophonic
- ❑ Deep Performer: Violin, Piano (EPS only)

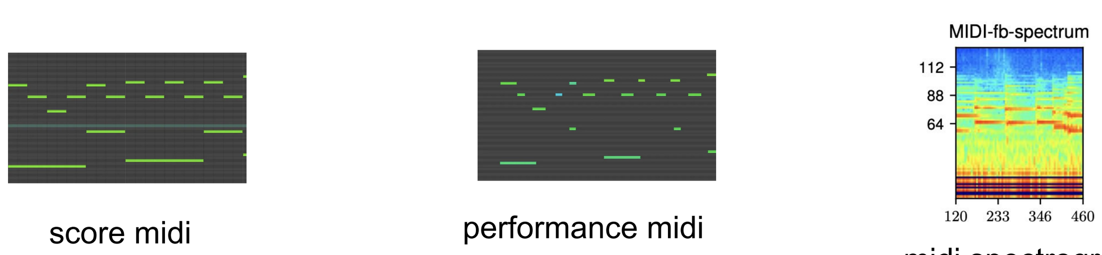
Propose Method



MIDI to MIDI (M2M)-102M MIDI to Audio (M2A)-12M

Score MIDI to Audio (S2A)-102M

Proposed System (Two-Stage) **Baseline (Single-Stage)**



score midi performance midi midi spectrogram

Fine-tuned M2A

- ❑ **ATEPP(>700h, >10000 recordings):transcribed** MIDIs
- ❑ **Maestre (~200h, ~1300 recordings): recorded** MIDIs
- ❑ Bridge the **knowledge gap** between to improve the synthesis quality

Baseline

- ❑ **Similar to M2A model, input replaced by score MIDI**
- ❑ We used the fine-tuned M2A to initialise the training, and further trained the baseline model with audio and score MIDI pairs

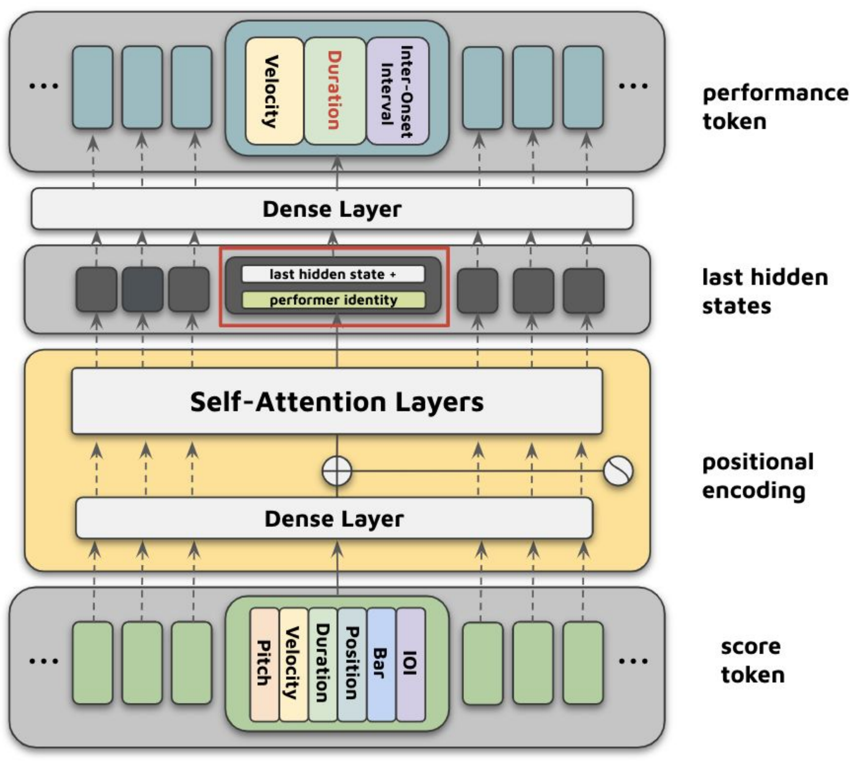


TABLE I
VOCABULARY SIZE OF THE TOKENISED NOTE-LEVEL FEATURES

Feature	Pitch	Vel.	Dur.	IOI	Pos.	Bar
Size	92	68	1156	772	388	3004

M2M Model (Improvements)

- ❑ **Reduced Vocabulary and Model Parameter Size:**
  - ❑ Adapted Octuple tokenisation method,
  - ❑ Used lower beat resolution to reduced vocabulary size
  - ❑ Segmented MIDI into 256-note sequences
- ❑ **Enhanced Pianist Identity Representation:** Identity embeddings were summed with the last hidden state
- ❑ **Improved Performance Generation:**
  - ❑ Predicted actual note durations instead of deviations.
  - ❑ Used probabilistic loss and sampling techniques to enhance output variety.

Objective Metrics and Subjective Evaluation

TABLE II
PERFORMANCE METRICS FOR THE M2M MODEL: MEANS AND 95% CONFIDENCE INTERVALS

	Performance-wise			*Segment-wise*		
Feature	**KLD ↓**	**Correlation ↑**	**DTWD ↓**	**KLD ↓**	**Correlation ↑**	**DTWD ↓**
Velocity	0.018 ± 0.001	0.831 ± 0.017	0.063 ± 0.002	0.016 ± 0.001	0.665 ± 0.011	0.064 ± 0.001
Inter-Onset Interval	< 0.001	0.990 ± 0.003	0.010 ± 0.001	< 0.001	0.932 ± 0.006	0.009 ± 0.001
Duration	0.187 ± 0.010	0.753 ± 0.017	0.026 ± 0.003	0.184 ± 0.005	0.668 ± 0.012	0.023 ± 0.001

- ❑ Effectively reconstructs IOI and velocity; duration prediction needs improvement.
- ❑ Segmenting performances into 256-note sequences had no negative effect on generation quality.
- ❑ Lower correlation in segments suggests full-performance assessments may overlook local inconsistencies.

TABLE III
PERFORMANCE METRICS FOR THE M2A MODEL AND BASELINE

Systems	Chroma ↓	Spectrogram ↓
Pianoteq	**0.487±0.008**	0.294±0.013
Baseline	0.624±0.027	0.284±0.013
M2A [18]	0.539±0.021	0.318±0.013
Fine-tuned M2A (**ours**)	0.522±0.018	**0.262±0.009**

- ❑ Fine-tuning the M2A model improved synthesis, reducing MIDI spectrogram distortion and enhancing ambient sound.
- ❑ The fine-tuned M2A model showed lower key accuracy than Pianoteq, but better captured nuanced performance details.
- ❑ The baseline model struggled with pitch reconstruction.

Listening Tests

TABLE IV
MEAN OPINION SCORES (MOS) FOR EXPRESSIVENESS AND QUALITY IN THE EVALUATED SYSTEMS FROM THE TWO LISTENING TESTS

Test A: *Evaluating the M2M Model: **Expressiveness***	
Systems (Midi Source + Synthesiser)	**MOS**
S1. Groundtruth (GT.) + Pianoteq (Ref.)	8.58 ± 0.41
S2. M2M Output (**ours**) + Pianoteq	6.69 ± 0.49
S3. M2M Output + Fine-tuned M2A (**ours**)	3.87 ± 0.57
S4. Score + Pianoteq	5.07 ± 0.58
S5. Score + Baseline	1.54 ± 0.38
Test B: *Evaluating the M2A Model: **Quality***	
Systems (Midi Source + Synthesiser)	**MOS**
S0. Human Performance Recording (Ref.)	7.19 ± 0.39
S1. Groundtruth + Pianoteq	7.41 ± 0.40
S6. Groundtruth + M2A [18]	6.29 ± 0.49
S7. Groundtruth + Fine-tuned M2A (**ours**)	5.96 ± 0.47
S3. M2M Output + Fine-tuned M2A (**ours**)	5.27 ± 0.52
S5. Score + Baseline	5.12 ± 0.59

- ❑ M2M model (**S2**) is more expressive than scores (**S4**) but less than human performances (**S1**) and generalizes well to unseen compositions.
- ❑ Fine-tuned M2A model (**S7**) had lower ratings than the original M2A (**S6**), excelling in ambient sound but also reproducing noise and inaccuracies.
- ❑ The two-stage system (**S3**) outperformed the single-stage baseline (**S5**).

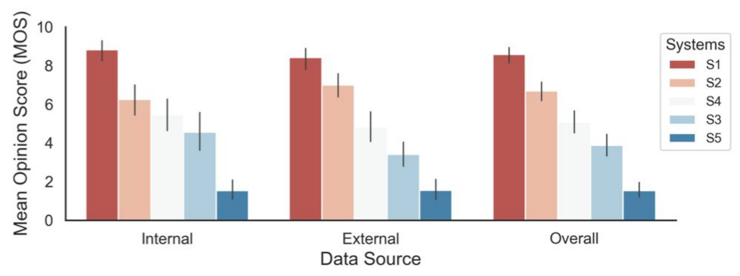


Fig. 3. MOS for systems listed in Table IV from Test A. Scores with respect to the internal and external compositions are presented.

Conclusion

- ❑ The two-stage system (M2M + fine-tuned M2A) enhances human-like expressiveness and preserves acoustic ambience, outperforming baseline models.
- ❑ Limitations include inconsistent acoustic ambience across full performances.
- ❑ Future work will focus on pedalling prediction, Chromagram loss, and improving performance across different environments and styles.

This work is supported by both the UKRI Centre for Doctoral Training in Artificial Intelligence and Music [grant number EP/S022694/1], and the National Institute of Informatics in Japan. J.Tang is a research student supported jointly by the China Scholarship Council [grant number 202008440382] and Queen Mary University of London. E. Cooper conducted this work while at NII, Japan and is currently employed by NICT, Japan.

UK Research and Innovation

Queen Mary University of London