Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset

Curtis Hawthorne et al., 2019

Rui Yoshinaga

October 4, 2021

Overview

Enabling Factorized Piano Music Modeling and Generation with the MAESTRO dataset

- Topic: Making Dataset, Music Generation
- Contributions:
  - MAESTRO dataset: Piano performance audio & aligned MIDI
  - Wave2Midi2Wave: Musical audio modeling
  - Achieved sota result on a transcription model

- Generated samples: `https://goo.gl/magenta/maestro-examples`
- MAESTRO dataset: `https://g.co/magenta/maestro-dataset`

Introduction

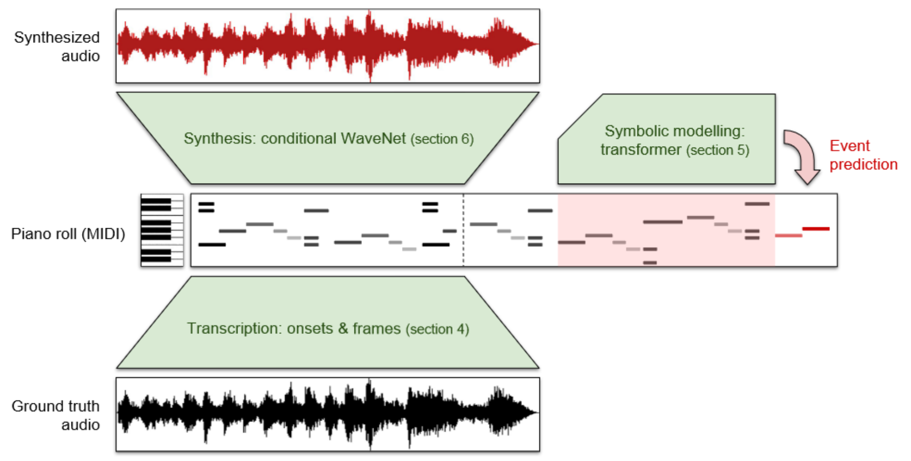
- Explicitly factorize the problem:

$$P(audio) = \mathbb{E}_{notes}\left[P(audio \mid notes)\right] \tag{1}$$

  which can be thought of as a generative model with notes

- Split the model into three separately trained modules:
  1. *Encoder*, $P(notes \mid audio)$:
     - Transcription model: Onsets and Frames [1]
     - Produce MIDI from raw audio
  2. *Prior*, $P(notes)$:
     - Language model: Music Transformer [2]
     - Generate new performances in MIDI format
  3. *Decoder*, $P(audio \mid notes)$:
     - Synthesis model: WaveNet [3]
     - Generate audio of the performances conditioned on MIDI

Wave2Midi2Wave


Synthesized audio
Synthesis: conditional WaveNet (section 6)
Symbolic modelling: transformer (section 5)
Event prediction
Piano roll (MIDI)
Transcription: onsets & frames (section 4)
Ground truth audio

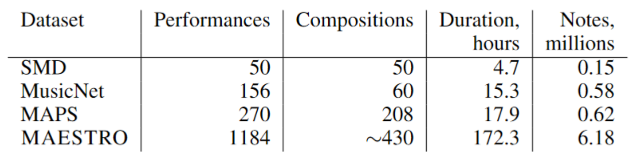
Wave2Midi2Wave system architecture

Dataset

MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization)

- Contains over a week of paired audio and MIDI recording
  - From 9 years of International Piano-e-Competition events
- Repertoire: Mostly classical (17th to early 20th century)

- Audio: CD quality or higher (44.1kHz–48kHz 16bit PCM stereo)
- MIDI: Includes key strike velocities, sustain pedal position
  - Audio and MIDI files are aligned with ≈ 3ms accuracy
  - Anotated with composer, title, year of performance



Dataset	Performances	Compositions	Duration, hours	Notes, millions
SMD	50	50	4.7	0.15
MusicNet	156	60	15.3	0.58
MAPS	270	208	17.9	0.62
MAESTRO	1184	∼430	172.3	6.18

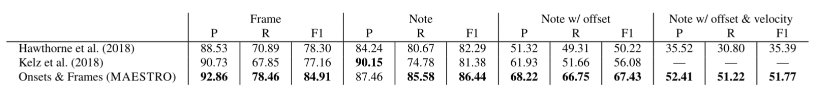
Piano Transcription (Wave2Midi)

Piano Transcription

- Model: Onsets and Frames [1] (with several modifications)
- Train the model on MAESTRO dataset
  - The best ways to get higher performance with the larger dataset were to make the model larger and simpler

Results (MAPS)

- Task: Piano transcription
- Test data: MAPS configuration 2 test dataset
- Achieved state of the art results for the MAPS dataset

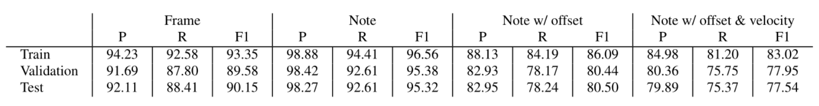


	Frame			Note			Note w/ offset			Note w/ offset & velocity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Hawthorne et al. (2018)	88.53	70.89	78.30	84.24	80.67	82.29	51.32	49.31	50.22	35.52	30.80	35.39
Kelz et al. (2018)	90.73	67.85	77.16	**90.15**	74.78	81.38	61.93	51.66	56.08	—	—	—
Onsets & Frames (MAESTRO)	**92.86**	**78.46**	**84.91**	87.46	**85.58**	**86.44**	**68.22**	**66.75**	**67.43**	**52.41**	**51.22**	**51.77**

Piano Transcription (Wave2Midi)

Results (MAESTRO)

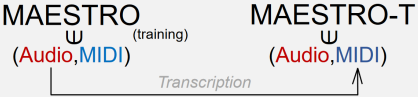
- Task: Piano transcription
- Apply the transcription model to the MAESTRO dataset
- Present results on the MAESTRO dataset as a new baseline score



	Frame			Note			Note w/ offset			Note w/ offset & velocity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Train	94.23	92.58	93.35	98.88	94.41	96.56	88.13	84.19	86.09	84.98	81.20	83.02
Validation	91.69	87.80	89.58	98.42	92.61	95.38	82.93	78.17	80.44	80.36	75.75	77.95
Test	92.11	88.41	90.15	98.27	92.61	95.32	82.95	78.24	80.50	79.89	75.37	77.54

MAESTRO-T

- Transcribe the audio in the MAESTRO training set
- Dataset with MAESTRO's audio and MIDI transcribed of it



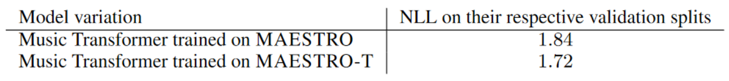
Music Transformer Training

Music Generation

- Model: Music Transformer [2]
- Format: Piano performance MIDI

Training

- Training Music Transformer on { MAESTRO, MAESTRO-T }
  - Data augmentation: transposition, time compression/stretching
- Model evaluation:



Model variation	NLL on their respective validation splits
Music Transformer trained on MAESTRO	1.84
Music Transformer trained on MAESTRO-T	1.72

Samples

- `https://goo.gl/magenta/maestro-examples`

Piano Synthesis (Midi2Wave)

Audio Synthesis

- Model: WaveNet [3] (with several modifications)
- Provide a MIDI sequence as conditioning information
  - cf. linguistic features in TTS

Training

1. Unconditional: trained only with the audio from MAESTRO
2. Ground: trained with audio/MIDI pairs from MAESTRO
3. Transcribed: trained with audio and MIDI from MAESTRO-T

Samples

- Model recreates non-piano subtleties of the recording
- Feed one-hot year vector for timbral shift
  - cf. speaker conditioning in TTS

Listening Tests

- Presented users with two 20-second clips,
  asked which clip sounded more like a real recording of piano playing
- Models (Audio/MIDI):
  - Ground Truth Recordings
  - WaveNet Unconditioned
  - WaveNet Ground/Test
  - WaveNet Transcribed/Test
  - WaveNet Transcribed/Transformer

Listening Tests

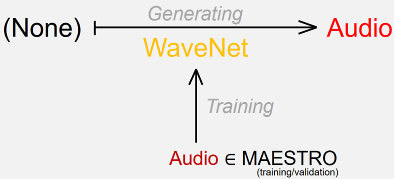
- Presented users with two 20-second clips,
  asked which clip sounded more like a real recording of piano playing

- Models (Audio/MIDI):
  - Ground Truth Recordings
  - WaveNet Unconditioned
  - WaveNet Ground/Test
  - WaveNet Transcribed/Test
  - WaveNet Transcribed/Transformer

Audio
⫙
MAESTRO
(validation)

Listening Tests

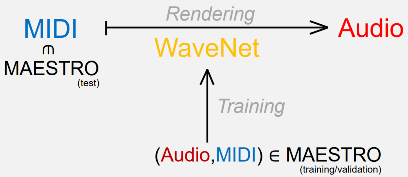
- Presented users with two 20-second clips,
  asked which clip sounded more like a real recording of piano playing
- Models (Audio/MIDI):
  - Ground Truth Recordings
  - WaveNet Unconditioned
  - WaveNet Ground/Test
  - WaveNet Transcribed/Test
  - WaveNet Transcribed/Transformer



Listening Tests

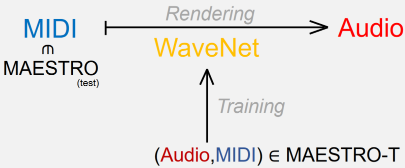
- Presented users with two 20-second clips,
  asked which clip sounded more like a real recording of piano playing

- Models (Audio/MIDI):
  - Ground Truth Recordings
  - WaveNet Unconditioned
  - WaveNet Ground/Test
  - WaveNet Transcribed/Test
  - WaveNet Transcribed/Transformer



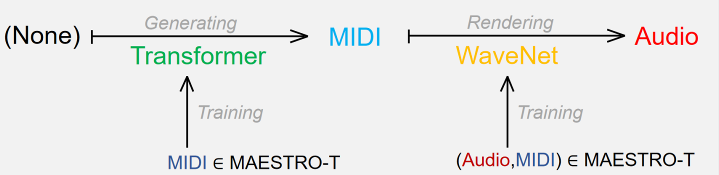
Listening Tests

- Presented users with two 20-second clips,
  asked which clip sounded more like a real recording of piano playing
- Models (Audio/MIDI):
  - Ground Truth Recordings
  - WaveNet Unconditioned
  - WaveNet Ground/Test
  - WaveNet Transcribed/Test
  - WaveNet Transcribed/Transformer



Listening Tests

- Presented users with two 20-second clips,
  asked which clip sounded more like a real recording of piano playing
- Models (Audio/MIDI):
  - Ground Truth Recordings
  - WaveNet Unconditioned
  - WaveNet Ground/Test
  - WaveNet Transcribed/Test
  - WaveNet Transcribed/Transformer



(None) Generating Transformer MIDI Rendering WaveNet Audio

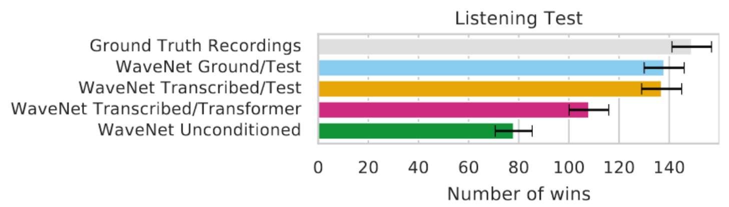
Training: MIDI ∈ MAESTRO-T

Training: (Audio,MIDI) ∈ MAESTRO-T

Listening Tests

- Presented users with two 20-second clips,
  asked which clip sounded more like a real recording of piano playing

- Results:
  - 640 ratings were collected



- Not a statistically significant difference
  - Ground Truth Recordings & WaveNet Ground/Test
  - Ground Truth Recordings & WaveNet Transcribed/Test

Conclusion

- MAESTRO dataset
    - A larger new dataset of piano performance
    - Used to train models in Wave2Midi2Wave
- Wave2Midi2Wave
    - Architecture combining existing models
    - Able to generate piano performances like real recordings
- Future work
    - Extend this approach to multiple simultaneous instruments

References

[1] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck.
Onsets and Frames: Dual-Objective Piano Transcription.
In Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, 2018, 2018.

[2] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck.
Music Transformer: Generating Music with Long-Term Structure.
In International Conference on Learning Representations, 2018.

[3] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu.
WaveNet: A Generative Model for Raw Audio.
arXiv preprint arXiv:1609.03499, 2016.

Dataset

Alignments

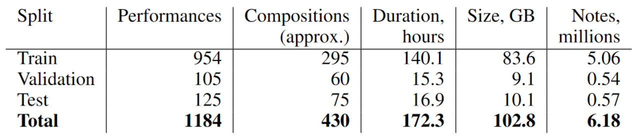
- MIDI files Recorded by Disklaviers themselves

  WAV audio Captured with conventional recording equipment

  Independent
- Approach: Globally minimizing the distance between CQT frames from real audio and synthesized MIDI

Dataset Spliting

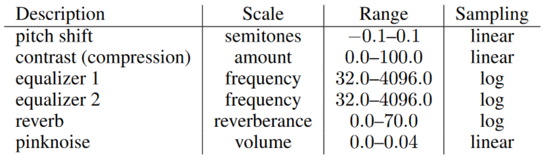
- Train/Validation/Test ≈ 80/10/10 %
- No composition should appear in more than one split
- The validation/test split should contain a variety of compositoins



Split	Performances	Compositions (approx.)	Duration, hours	Size, GB	Notes, millions
Train	954	295	140.1	83.6	5.06
Validation	105	60	15.3	9.1	0.54
Test	125	75	16.9	10.1	0.57
Total	**1184**	**430**	**172.3**	**102.8**	**6.18**

Data Augmentation on Transcription Model

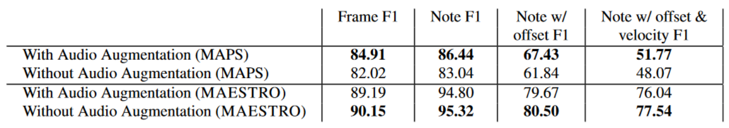
- Model: Onsets and Frames [1] (with several modifications)
- Training: MAESTRO training set (with audio augmentation)



Description	Scale	Range	Sampling
pitch shift	semitones	−0.1–0.1	linear
contrast (compression)	amount	0.0–100.0	linear
equalizer 1	frequency	32.0–4096.0	log
equalizer 2	frequency	32.0–4096.0	log
reverb	reverberance	0.0–70.0	log
pinknoise	volume	0.0–0.04	linear

- Evaluation: MAPS configuration 2 test set, MAESTRO test set

- Effects of audio augmentation:
  - was important on the MAPS dataset
  - made results slightly worse on MAESTRO dataset



	Frame F1	Note F1	Note w/ offset F1	Note w/ offset & velocity F1
With Audio Augmentation (MAPS)	**84.91**	**86.44**	**67.43**	**51.77**
Without Audio Augmentation (MAPS)	82.02	83.04	61.84	48.07
With Audio Augmentation (MAESTRO)	89.19	94.80	79.67	76.04
Without Audio Augmentation (MAESTRO)	**90.15**	**95.32**	**80.50**	**77.54**