icassp 2022 Singapore

Deep Performer: Score-to-Audio Music Performance Synthesis

Hao-Wen Dong[1,2] * Cong Zhou[1] Taylor Berg-Kirkpatrick[2] Julian McAuley[2]

[1] Dolby Laboratories [2] University of California San Diego

* Work done during an internship at Dolby

Dolby UC San Diego

Introduction

Music performance synthesis

- **Goal –** Synthesize a natural performance from a musical score

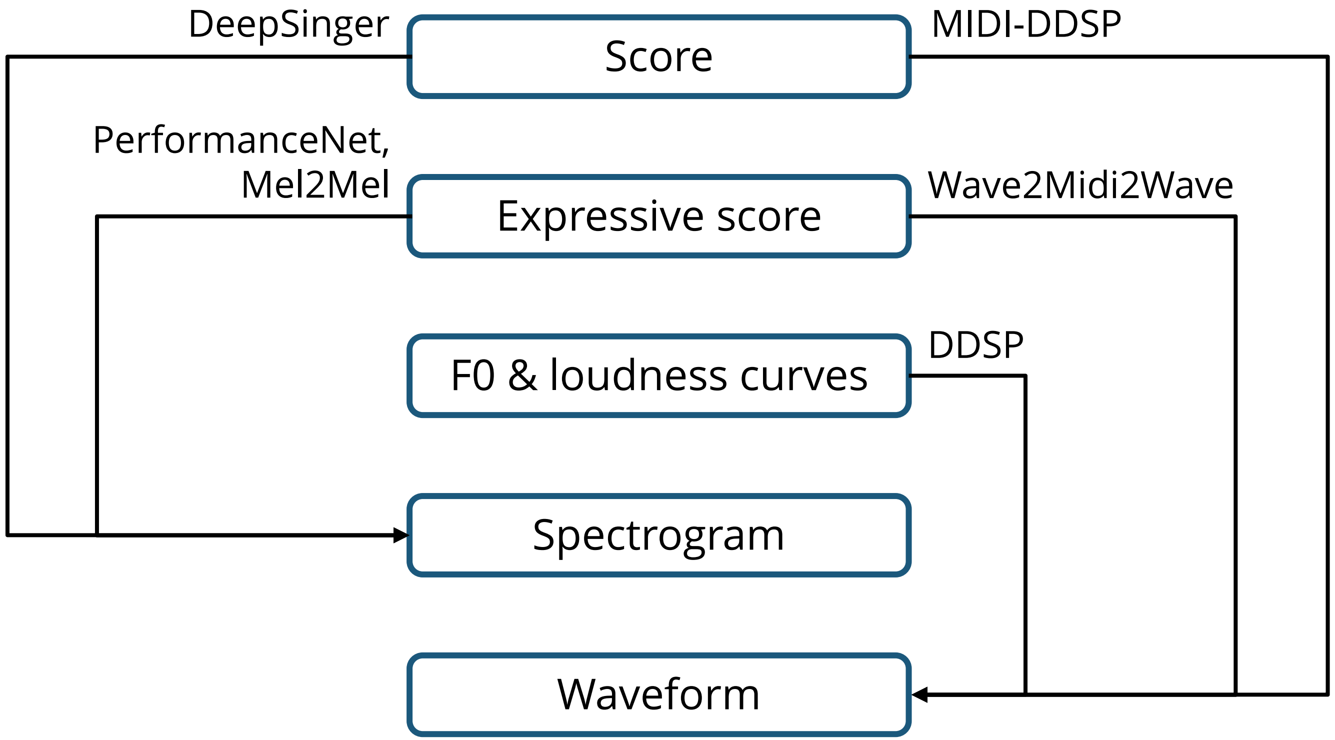
- Traditional synthesizers
  - Require costly samples (recordings of individual notes)
  - Do not model different playing styles and performative factors

- *Can we advance music synthesis with deep neural networks?*

Challenges

- Lack of paired training data
  - Hard to acquire paired data of musical scores and their recordings
  - Need to align the scores and the recordings
- Music often contains polyphony and long notes
  - Need to handle concurrent notes in the model
  - Need to provided fine-grained conditioning to the model

Prior work



Wang and Yang, “PerformanceNet: Score-to-Audio Music Generation with Multi-Band Convolutional Residual Network,” *Proc. AAAI*, 2019.
Hawthorne et al., “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset,” *Proc. ICLR*, 2019.
Kim et al., “Neural Music Synthesis for Flexible Timbre Control,” *Proc. ICASSP*, 2019.
Ren et al., “DeepSinger: Singing Voice Synthesis with Data Mined From the Web,” *Proc. KDD*, 2019.
Engel et al., “DDSP: Differentiable Digital Signal Processing,” *Proc. ICLR*, 2020.
Wu et al., “MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling,” *Proc. ICLR*, 2022.

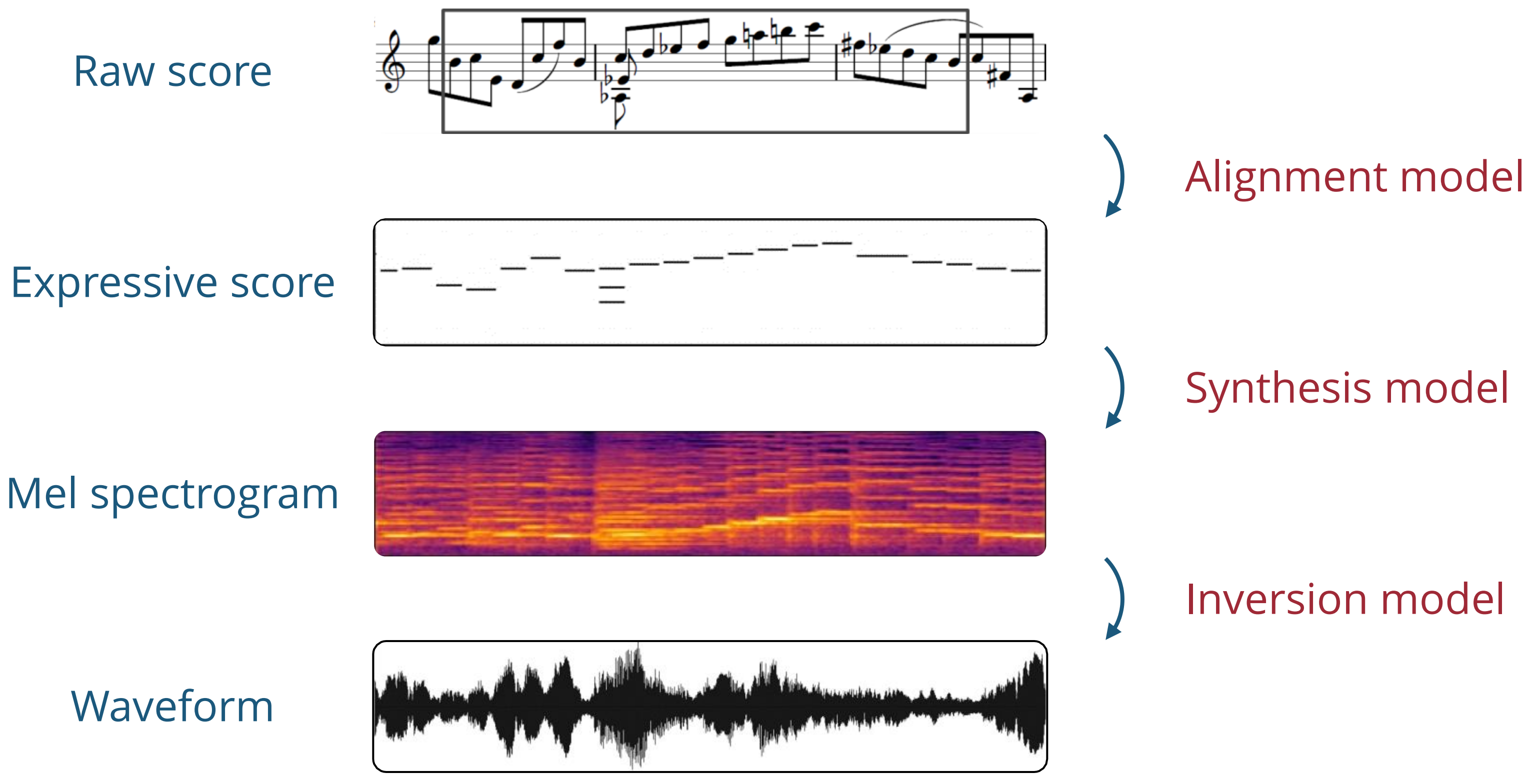
Prior work

Model	Unaligned inputs	Polyphonic inputs	Real recordings
PerformanceNet (Wang & Yang 2019)		✓	✓
Wave2Midi2Wave (Hawthorne et al. 2019)		✓	✓
Mel2Mel (Kim et al. 2019)		✓	
DeepSinger (Ren et al. 2019)	✓		✓
DDSP (Engel et al. 2020)			✓
MIDI-DDSP (Wu et al. 2022)	✓	✓	✓
Ours	✓	✓	✓

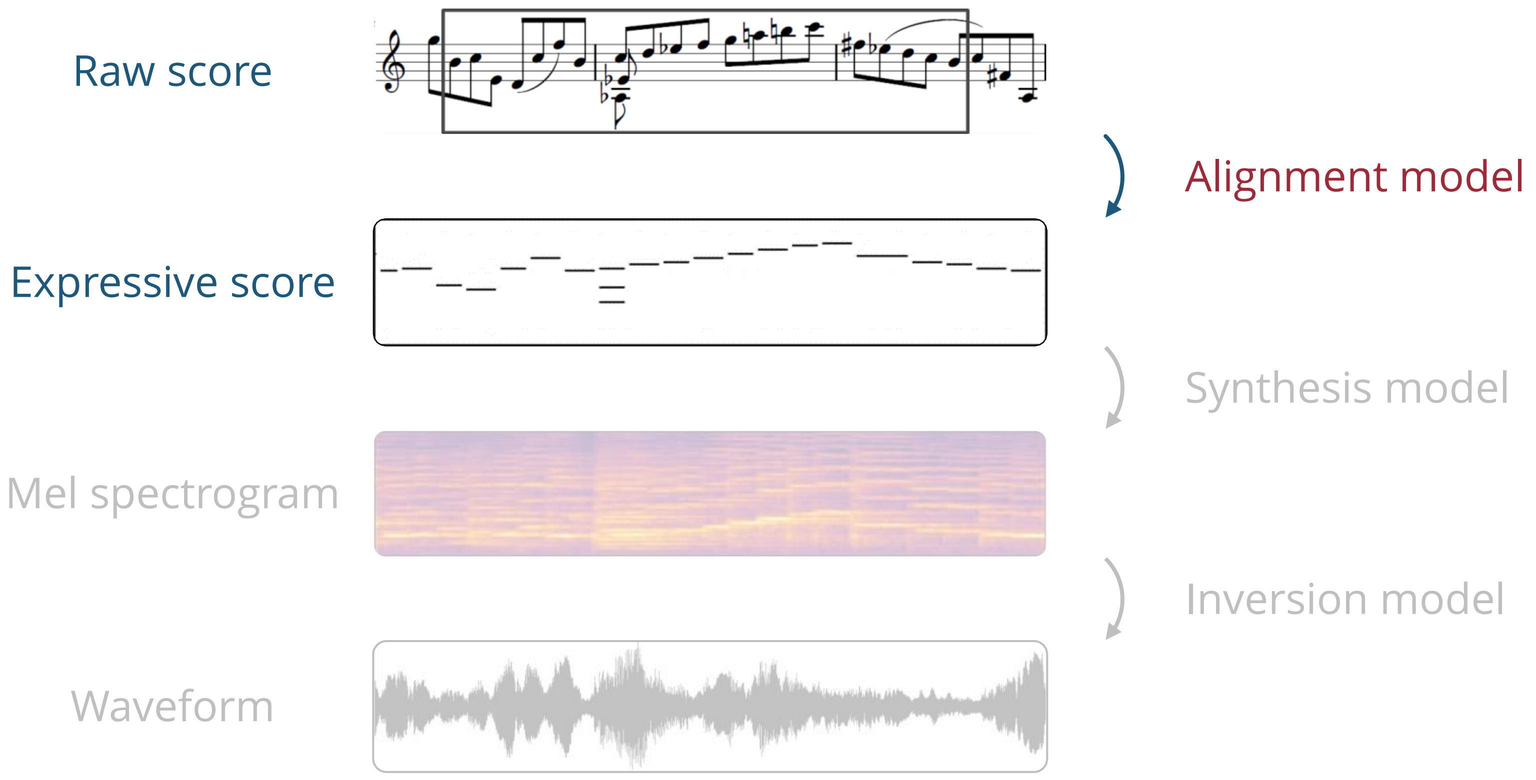
Wang and Yang, "PerformanceNet: Score-to-Audio Music Generation with Multi-Band Convolutional Residual Network," *Proc. AAAI*, 2019.
Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," *Proc. ICLR*, 2019.
Kim et al., "Neural Music Synthesis for Flexible Timbre Control," *Proc. ICASSP*, 2019.
Ren et al., "DeepSinger: Singing Voice Synthesis with Data Mined From the Web," *Proc. KDD*, 2019.
Engel et al., "DDSP: Differentiable Digital Signal Processing," *Proc. ICLR*, 2020.
Wu et al., "MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling," *Proc. ICLR*, 2022.

Model

Overview

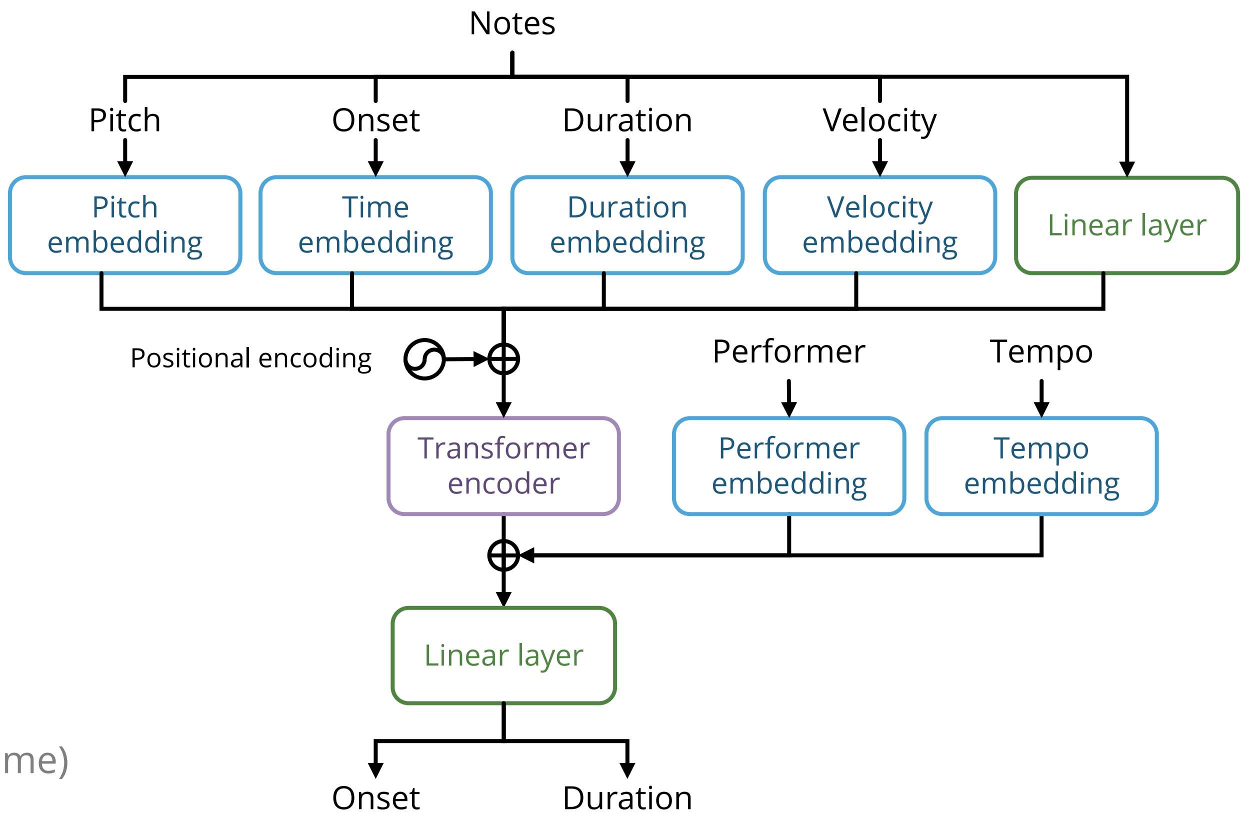

Raw score
Alignment model
Expressive score
Synthesis model
Mel spectrogram
Inversion model
Waveform

Alignment model

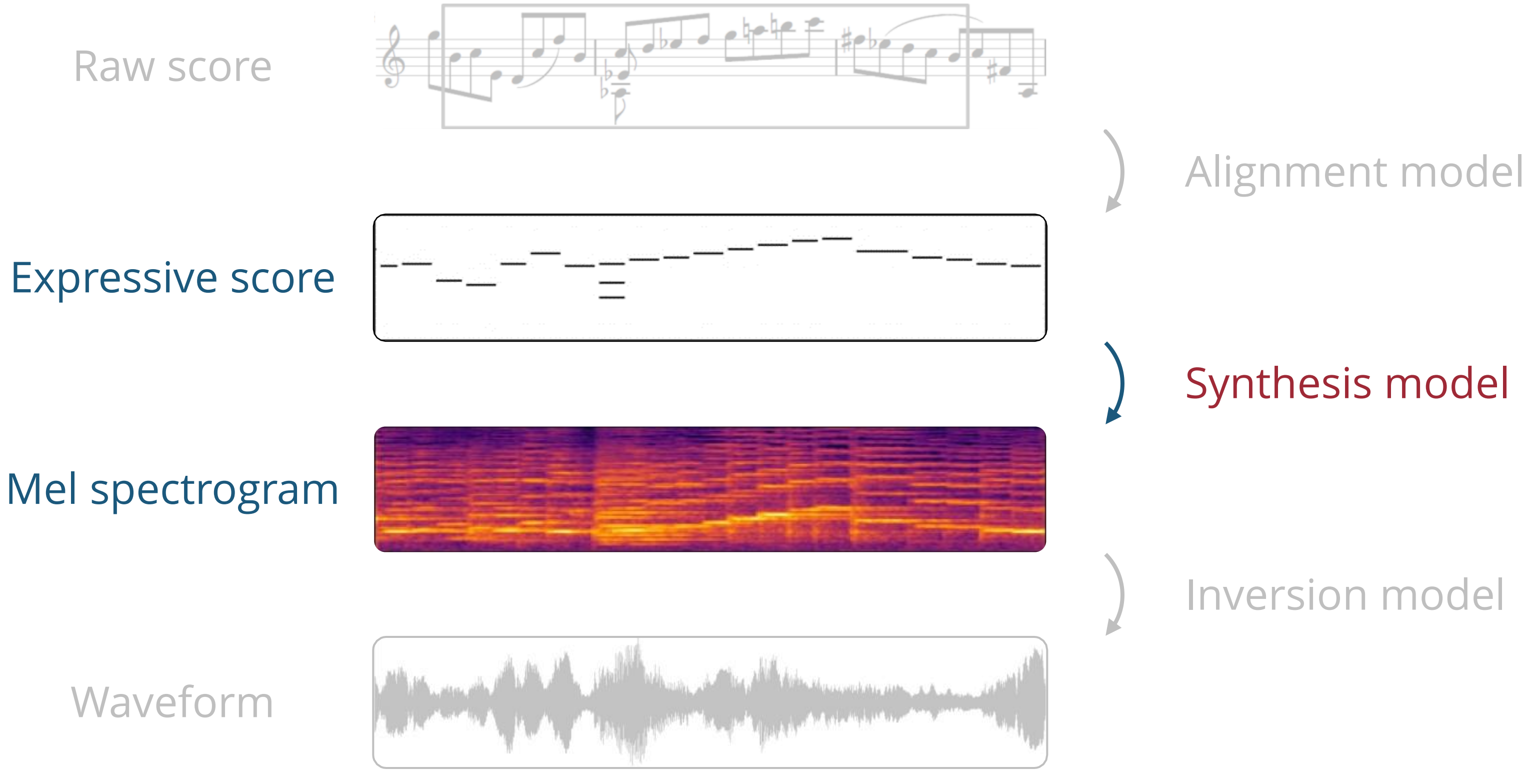


Alignment model

- A transformer encoder network
- **Input**
  - Note specified by its pitch, onset, duration and velocity
  - Performer ID
  - Tempo class
- **Output**
  - Expressive onset and duration (unit: frame)

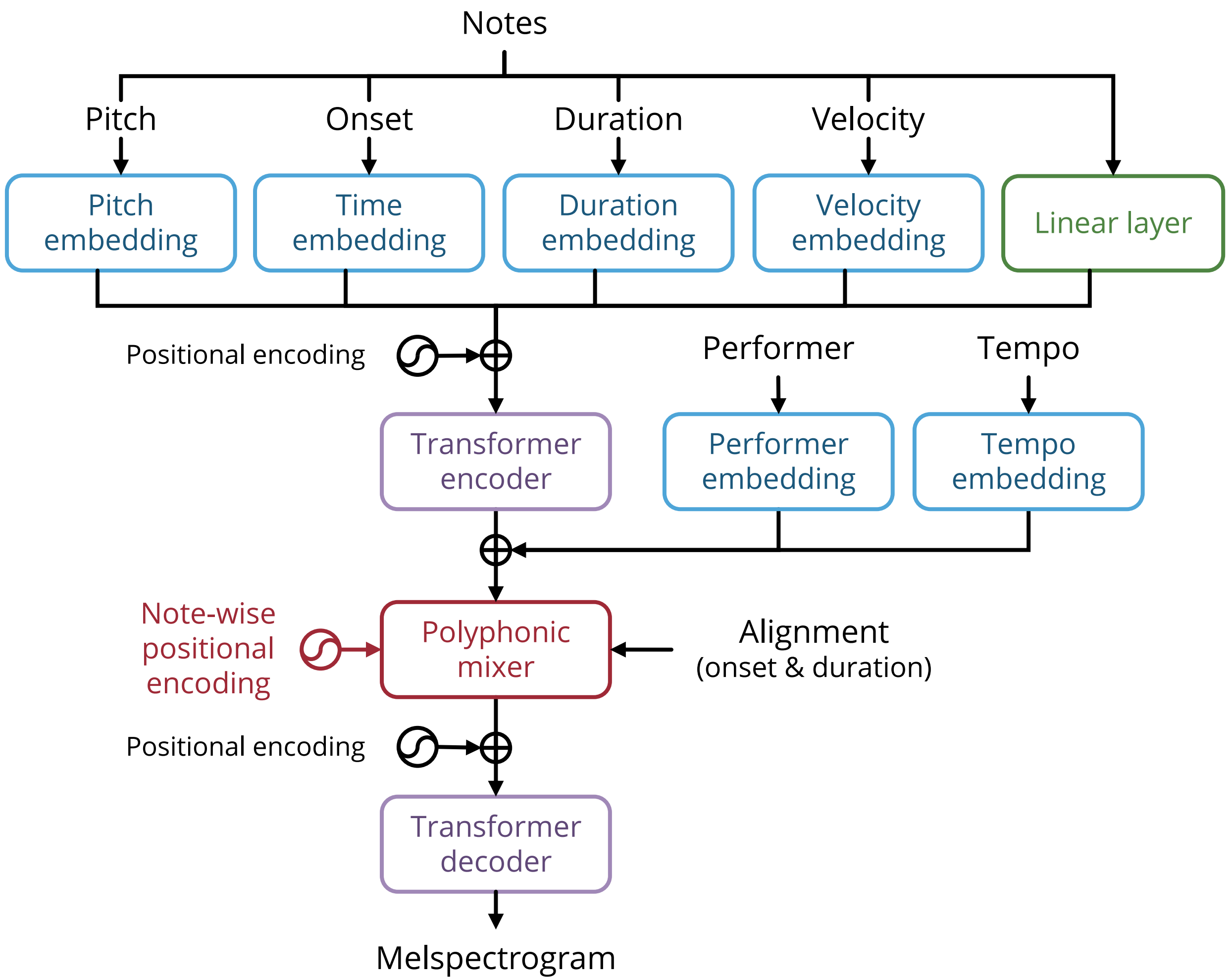


Synthesis model



Synthesis model

- A transformer network
- Based on FastSpeech (Ren et al. 2019)
- **Input**
  - Note specified by its pitch, onset, duration and velocity
  - Performer ID
  - Tempo class
  - Expressive onset and duration
- **Output**
  - Melspectrogram frames



Ren et al., "FastSpeech: Fast, robust and controllable text to speech," *Proc. NeurIPS*, 2019.

Proposed mechanisms

- Polyphonic mixer

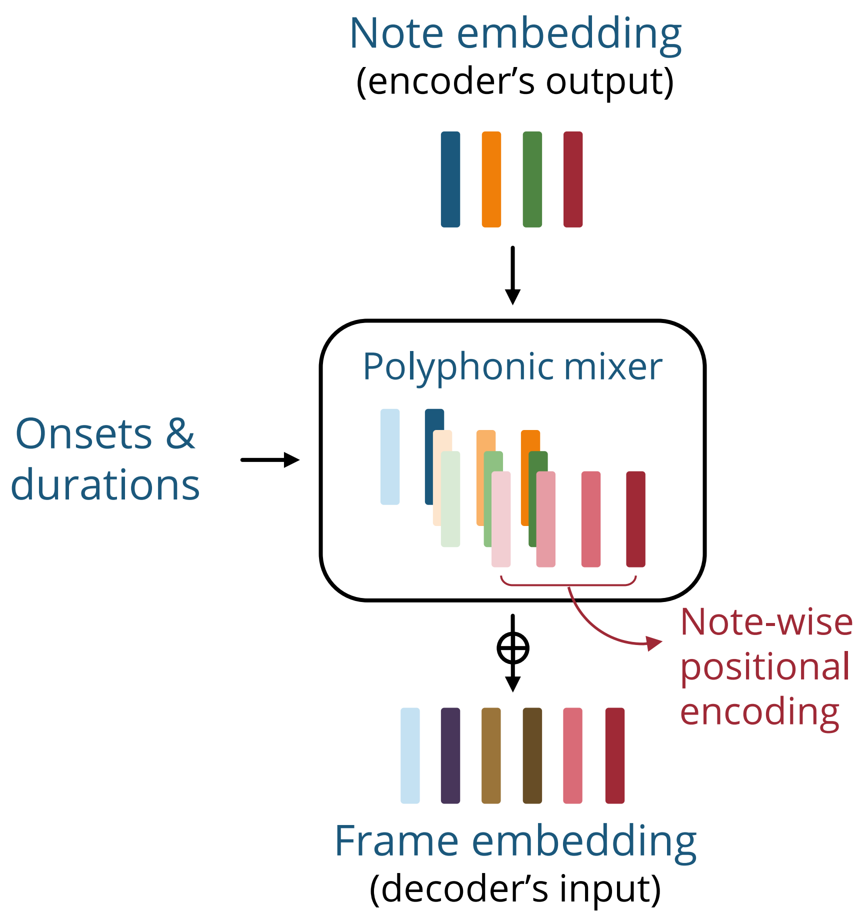
  Extend the state expansion mechanism to handle polyphonic inputs

- Note-wise positional encoding

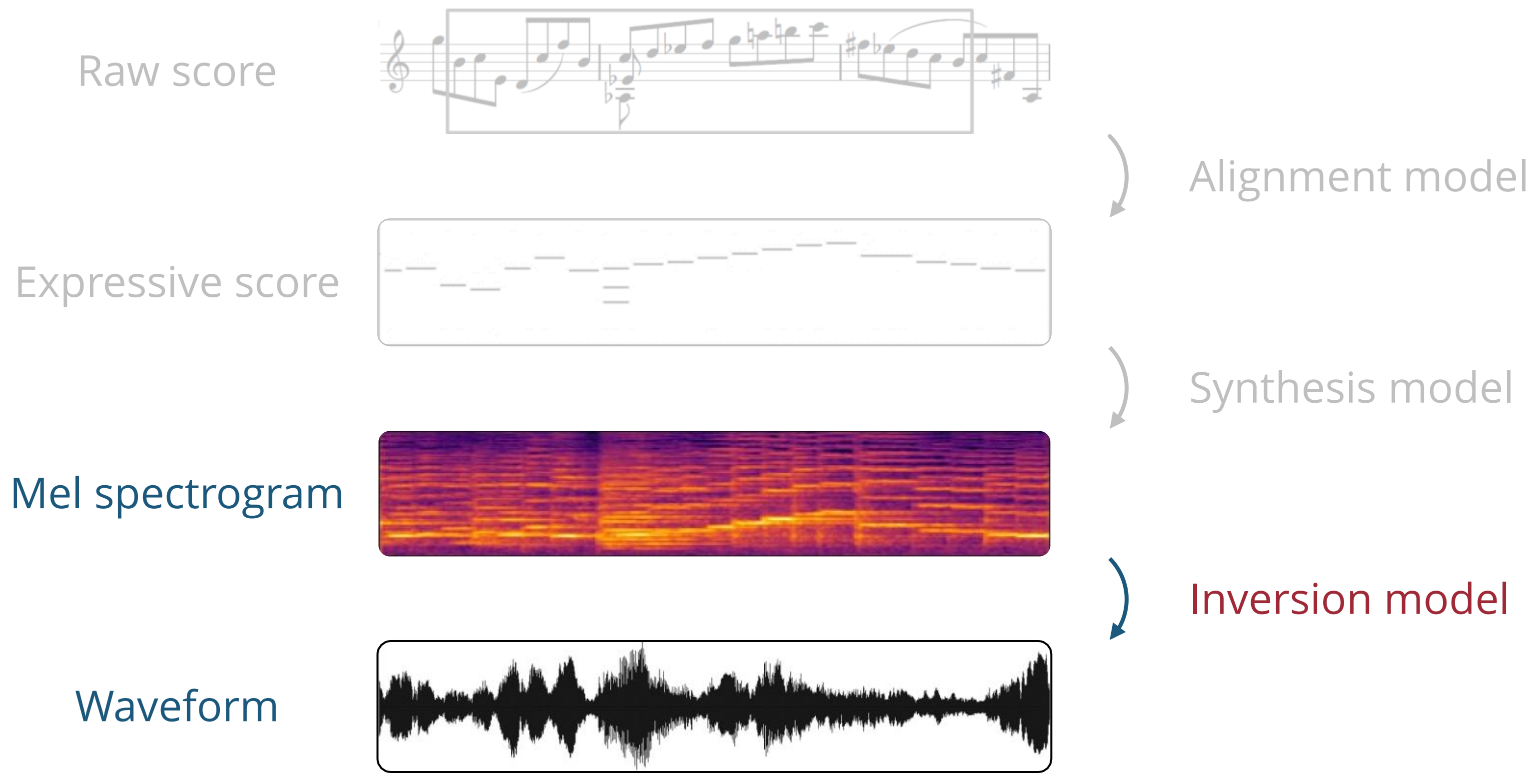
  Provide positional information within each note for a fine-grained conditioning

$$\mathbf{v}_{frame} = (1 + p\mathbf{w}) \odot \mathbf{v}_{note}$$

frame embedding learnable weight note embedding



Inversion model



Inversion model

- Hifi-GAN model (Kong et al. 2020)
  - Based on generative adversarial networks (GANs)

Kong et al., "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Proc. NeurIPS*, 2020.

Data

Bach Violin Dataset

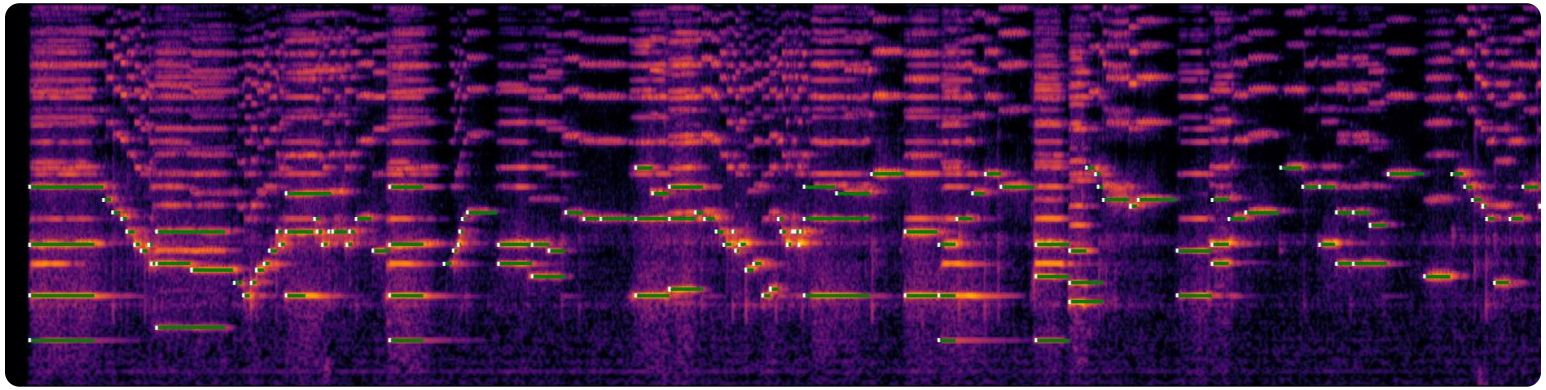
- Bach's sonatas and partitas for solo violin (BWV 1001–1006)
- 6.7 hours, 17 violinists

Dataset available at
salu133445.github.io/bach-violin-dataset/

Alignment derivation

1. Synthesize the scores using FluidSynth (a free software synthesizer)
2. Run dynamic time warping on the spectrograms (of the recording & synthesized audio)

Alignment result



Source code available at
github.com/salu133445/bach-violin-dataset

Experiments & Results

Implementation details

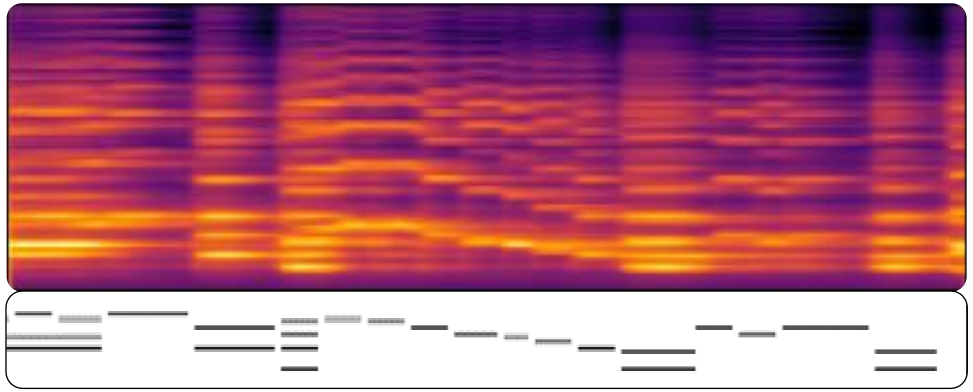
- **Audio**
  - mono, 16 kHz
- **Melspectrogram**
  - 80 Mel bands, STFT filter length: 1024, hop length: 256, window size: 1024
- **Alignment model**
  - 3 encoder layers (128 hidden neurons, 2 attention heads, 256 FFN hidden neurons)
- **Synthesis model**
  - 3 encoder layers, 6 decoder layers (128 hidden neurons, 2 attention heads, 512 FFN hidden neurons)
- **Training**
  - Adam optimizer (Kingma & Ba 2015)

Kingma and Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.

Demo

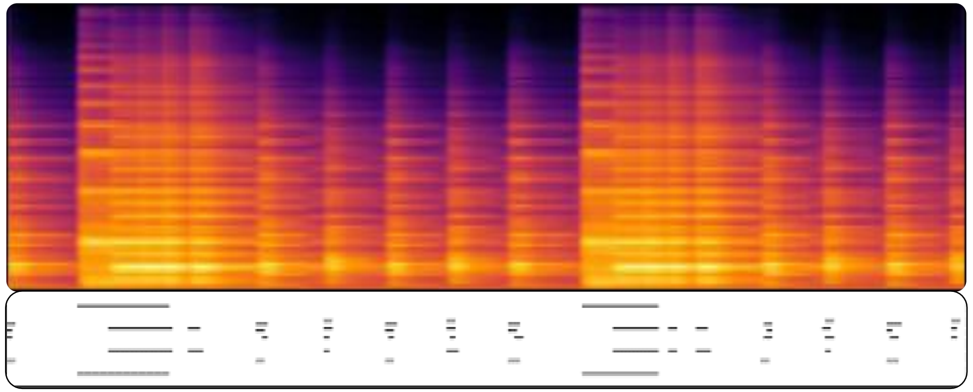
Violin

(trained on Bach Violin Dataset)



Piano

(trained on MAESTRO Dataset)

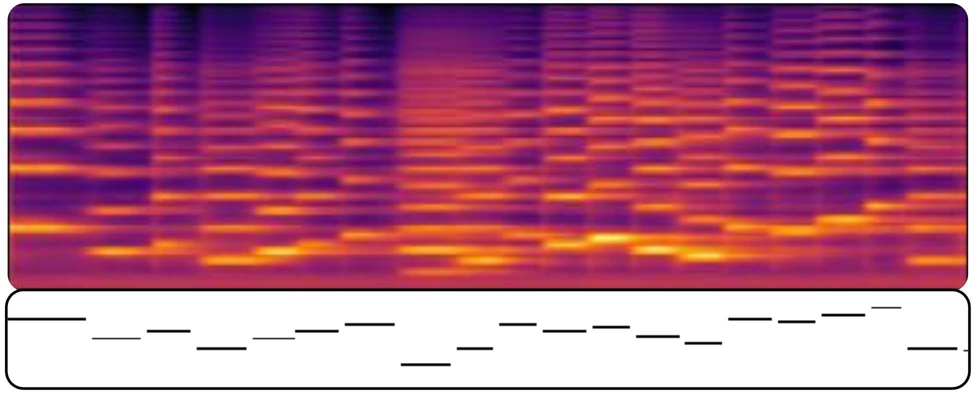


More samples available at
salu133445.github.io/deepperformer/

Comparisons to baseline

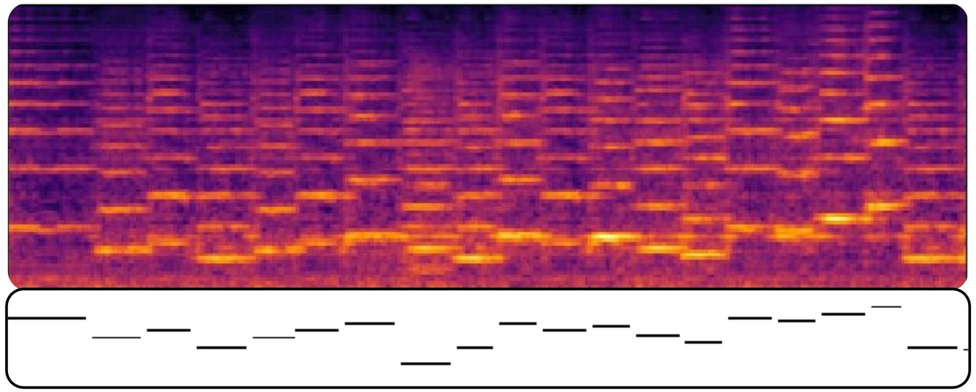
Deep Performer

(ours)



Hifi-GAN baseline

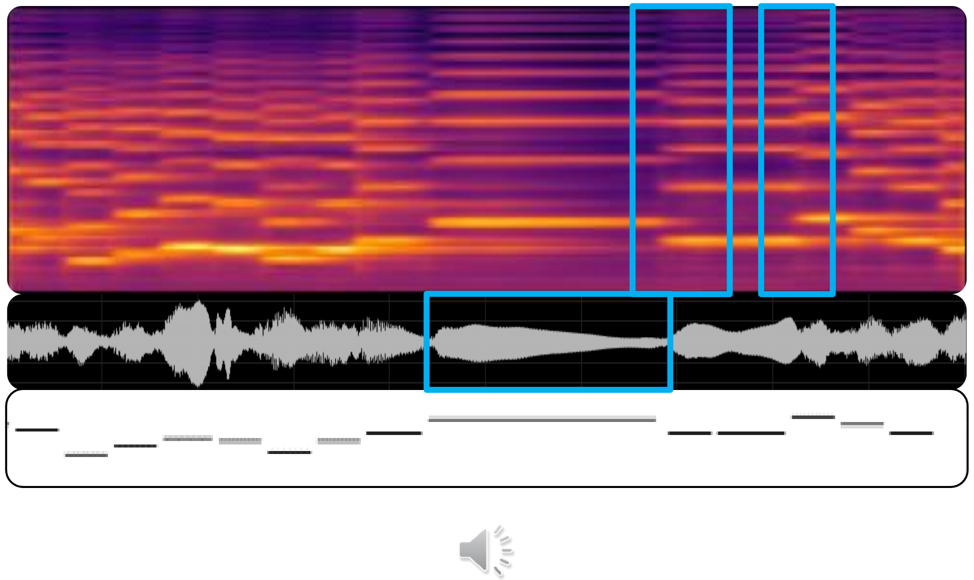
(piano roll conditioned)



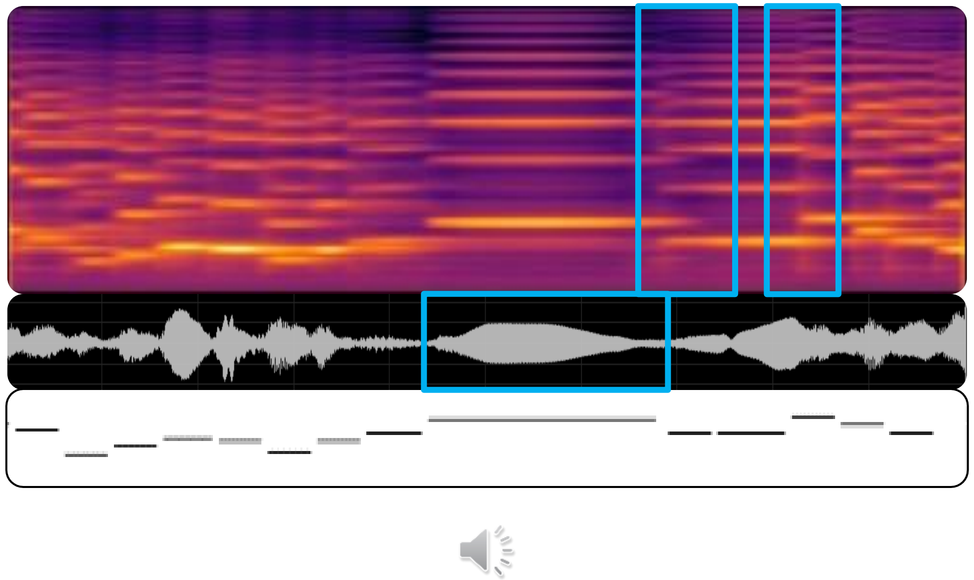
More samples available at
salu133445.github.io/deepperformer/

Note-wise positional encoding

With note-wise positional encoding



Without note-wise positional encoding



More samples available at
salu133445.github.io/deepperformer/

Subjective listening test

Model	Violin	Piano
Hifi-GAN baseline	2.57 ± 0.22	1.49 ± 0.17
Deep Performer (ours)	2.58 ± 0.21	2.17 ± 0.24
- w/o note-wise positional encoding	2.61 ± 0.23	2.37 ± 0.23
- w/o performer embedding	2.01 ± 0.25	2.26 ± 0.25
- w/o encoder (using piano-roll inputs)	2.22 ± 0.18	1.43 ± 0.16

(mean opinion scores reported)

Future Work

Modeling expressions & playing styles

Musical expressions

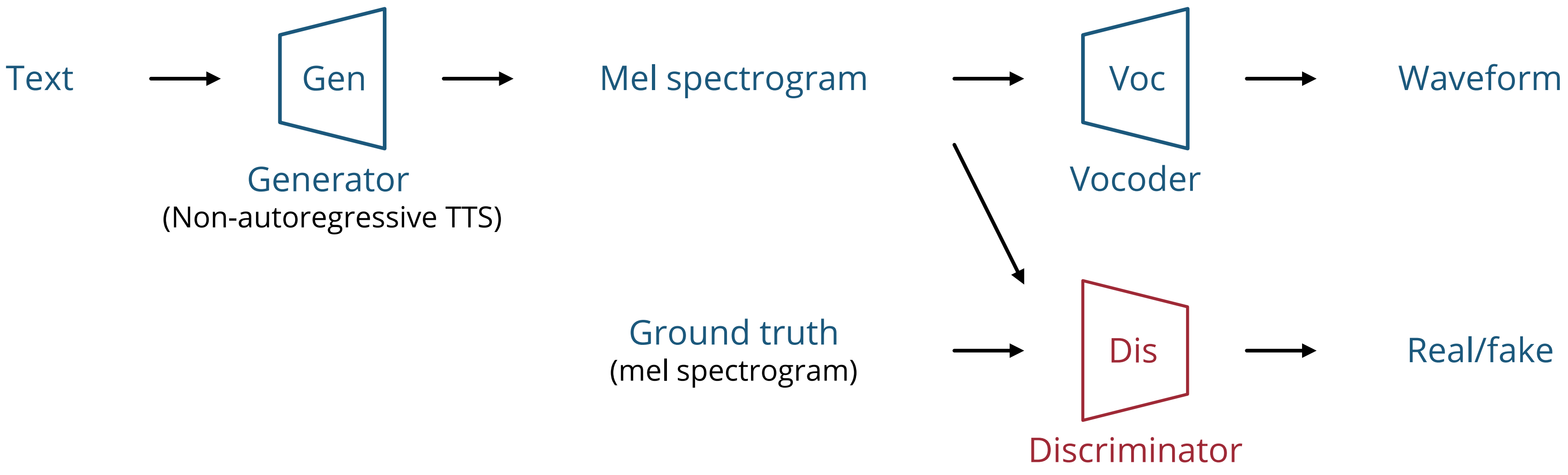
Dynamic, tempo, phrasing, articulation, etc.

Playing styles

Various musical interpretations of the same piece

Incorporating adversarial losses

- Improve the sharpness of the synthesized audio using adversarial losses
  - Promising results in speech synthesis (Yang et al. 2021)



Yang et al., "GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis," *Proc. INTERSPEECH*, 2021.

Conclusion

Conclusion

- Presented a new three-stage system for music performance synthesis
- Proposed two mechanisms for a transformer model
  - Polyphonic mixer for handling polyphonic inputs
  - Note-wise positional encoding for providing a fine-grained conditioning
- Showed the effectiveness of the proposed model
  - Outperforms the baseline on the piano dataset
  - Achieve competitive quality on the violin dataset

Thank you!

Learn more at salu133445.github.io/deepperformer/