Deep Performer: Score-to-Audio Music Performance Synthesis

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

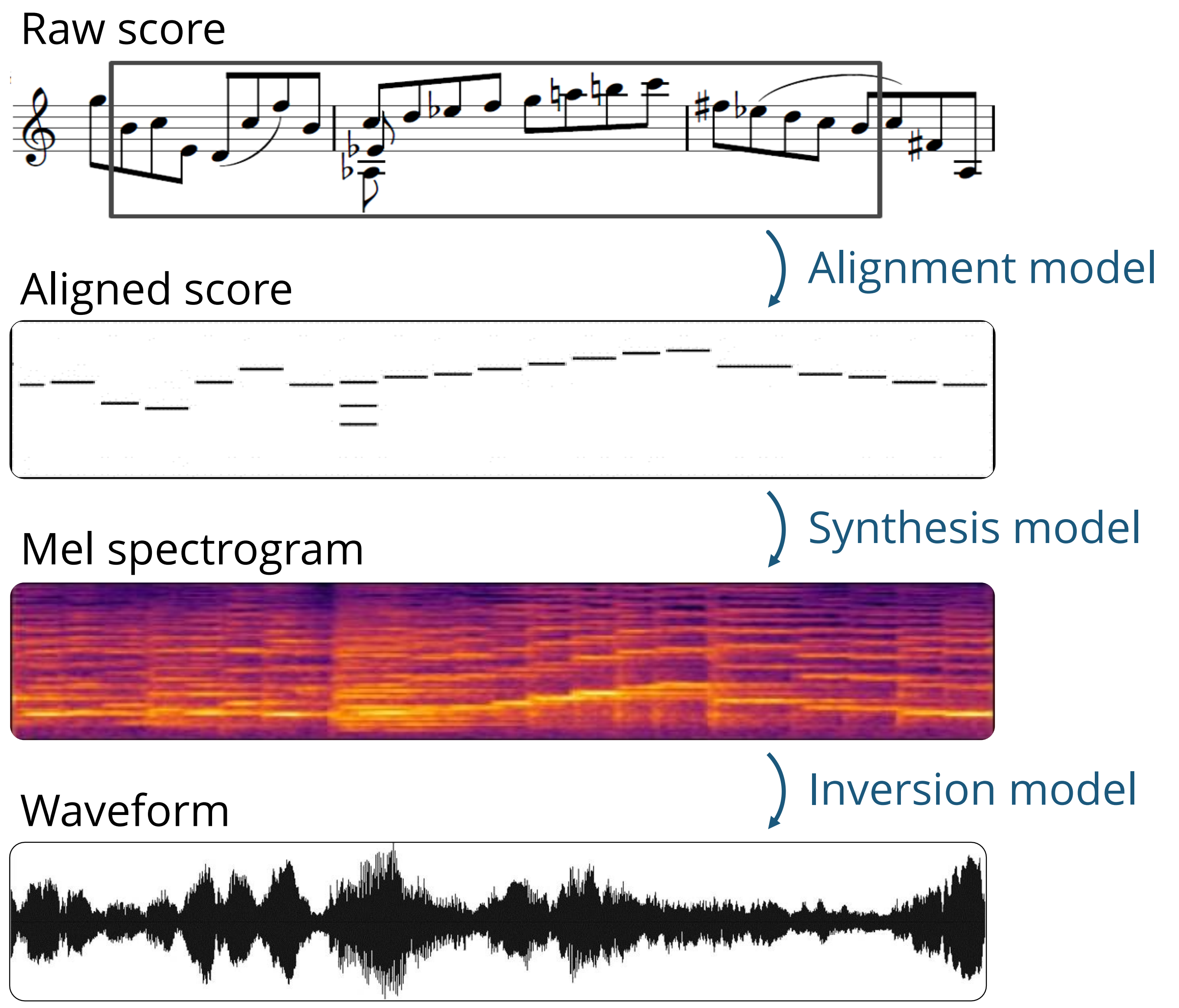
[1] Dolby Laboratories [2] University of California San Diego

* Work done during an internship at Dolby

Introduction

Music performance synthesis aims to synthesize a musical score into a natural performance. In this paper, we borrow recent advances in text-to-speech synthesis and present the Deep Performer—a novel system for score-to-audio music performance synthesis.

Overview



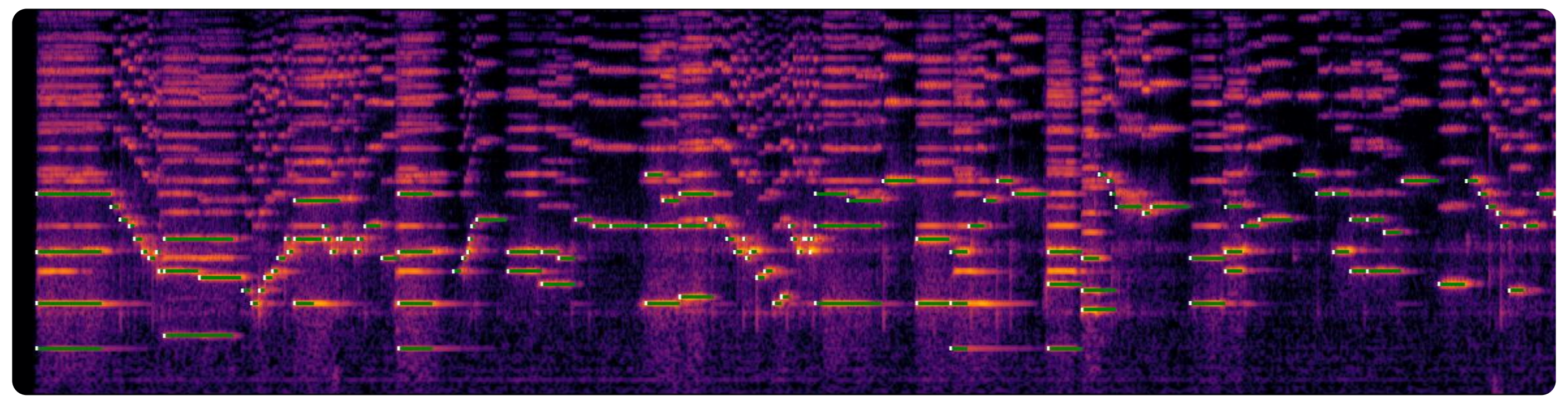
Data

Bach Violin Dataset

- Bach's sonatas and partitas for solo violin
- 6.7 hours, 17 violinists

Alignment derivation

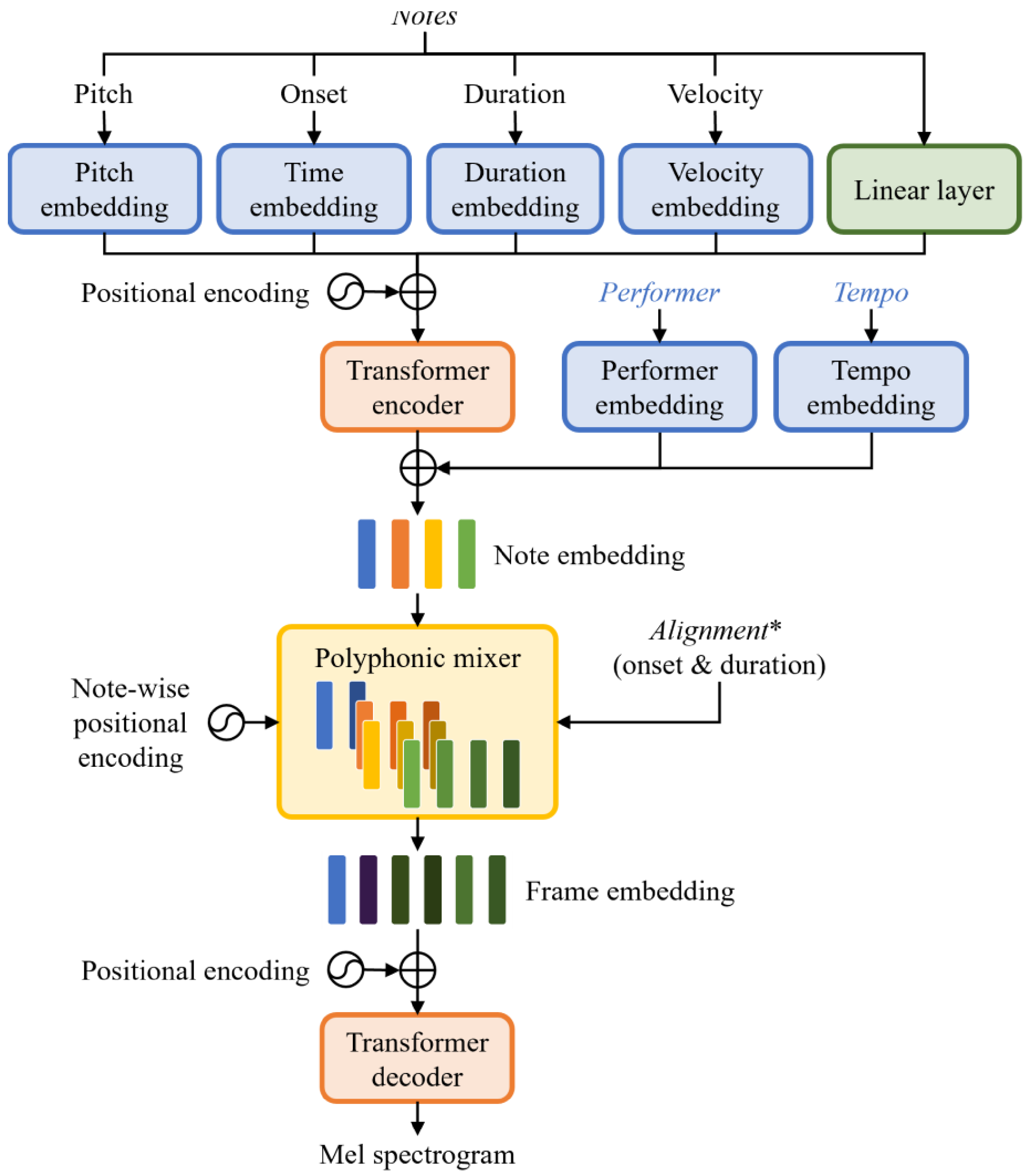
1. Synthesize the scores using FluidSynth
2. Run dynamic time warping on the spectrograms of the recording and synthesized audio



Model

Unlike speech, music often contains polyphony and long notes. Hence, we propose two new techniques for a transformer encoder-decoder model:

- The polyphonic mixer for handling polyphonic inputs
- The note-wise positional encoding for providing a fine-grained conditioning



Subjective Listening Test

We achieve competitive quality against the baseline model, a conditional generative audio model, in terms of pitch accuracy, timbre and noise level. Moreover, our proposed model significantly outperforms the baseline on an existing piano dataset in overall quality.

Audio samples can be found at
salu133445.github.io/deepperformer/.

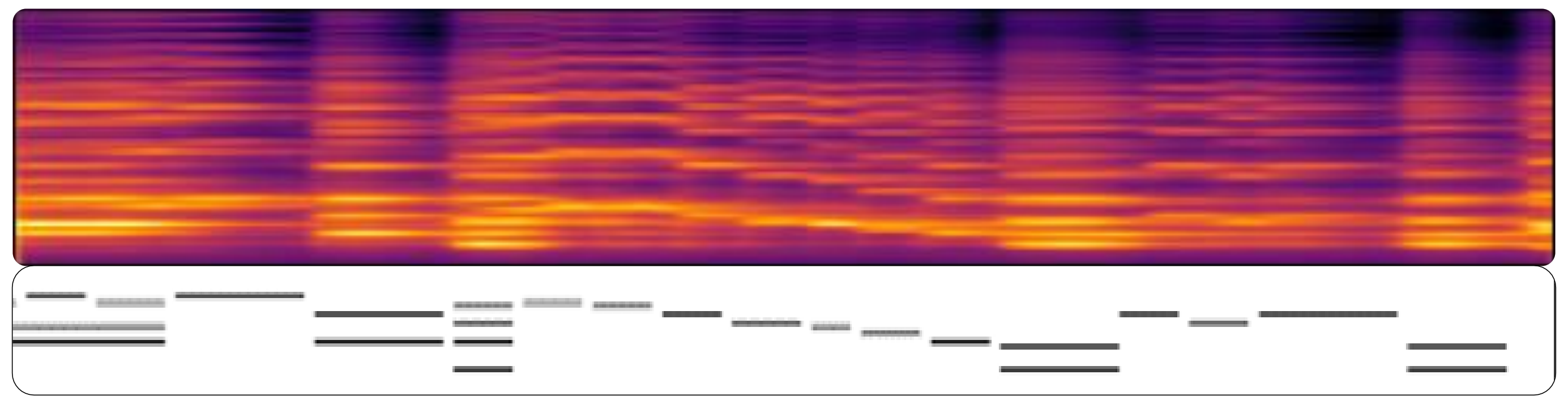
Listening test results (mean opinion scores reported)

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

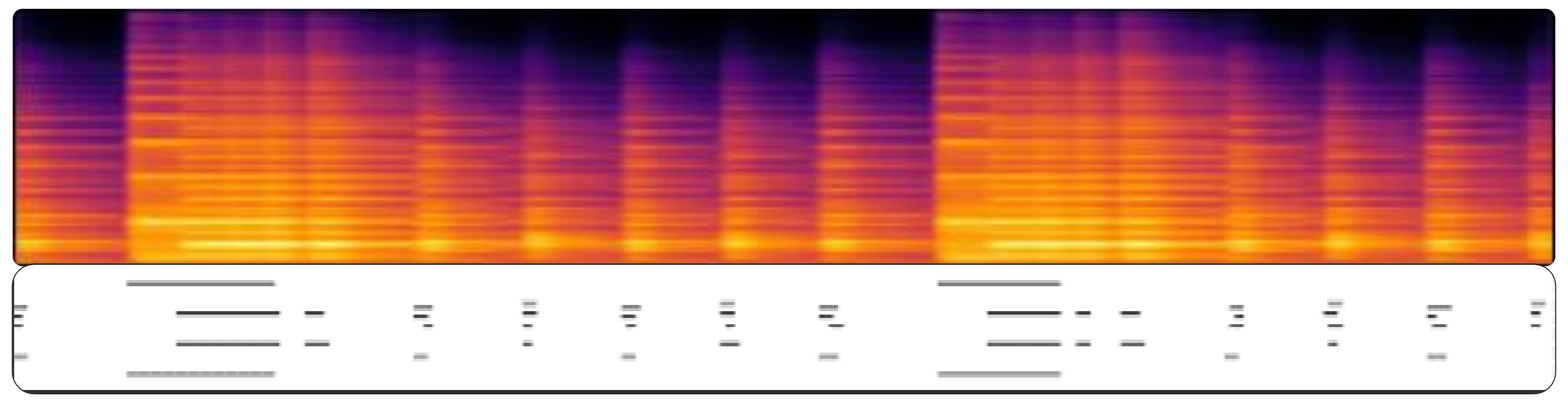
Results

Our proposed model can synthesize music with clear polyphony and harmonic structures.

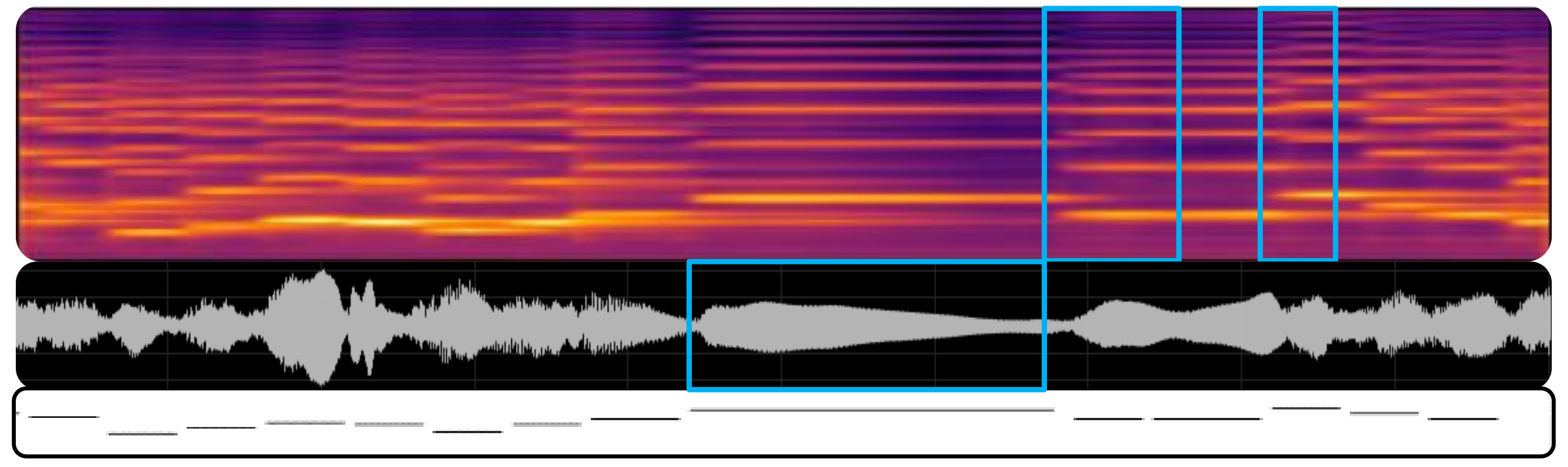
Violin example (on the Bach Violin Dataset)



Piano example (on the MAESTRO Dataset)



With note-wise positional encoding



Without note-wise positional encoding