Rubato

Transcribing Piano Music with Timestamps

Nazif Can Tamer, Victoria Ebert,
Guang Yang, Noah A. Smith

Contact: nctamer@cs.washington.edu

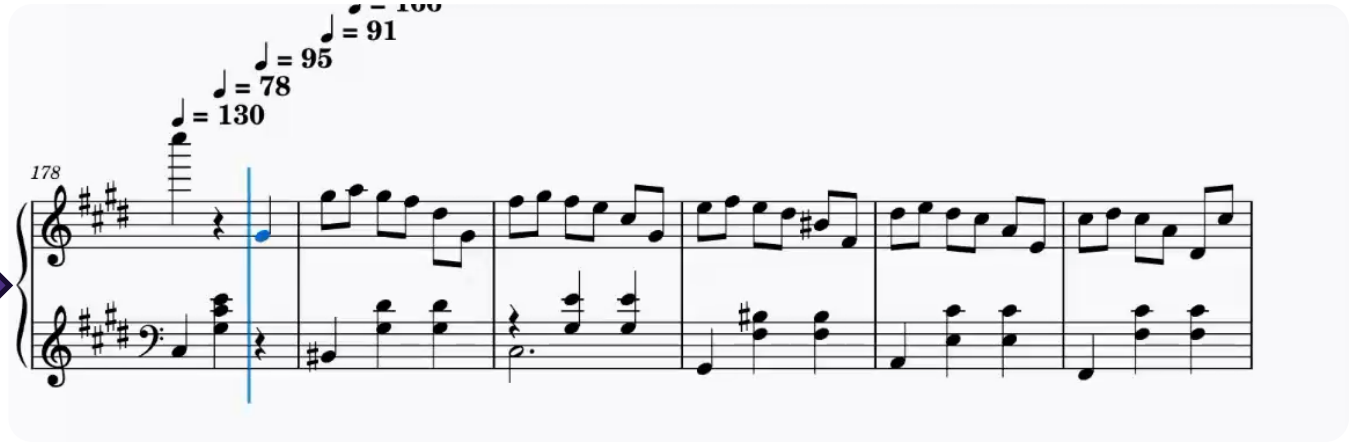
UNIVERSITY of WASHINGTON

The first end-to-end music transcription with…

- ✓ Score semantics
- ✓ Note grounding
- ✓ Beat grounding

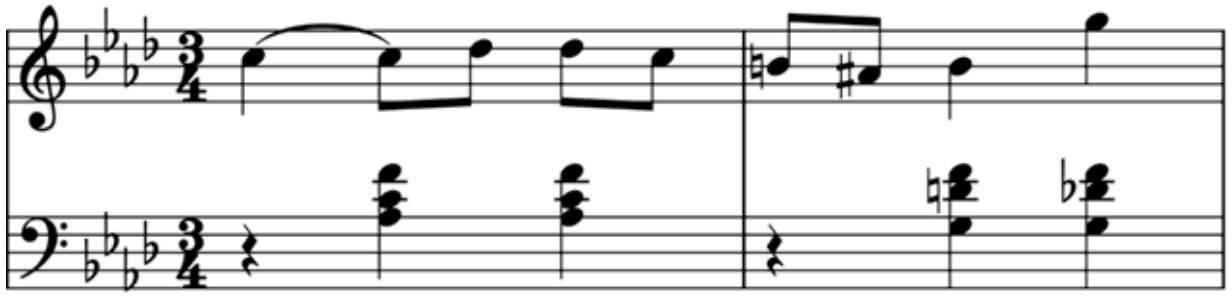


Rubato

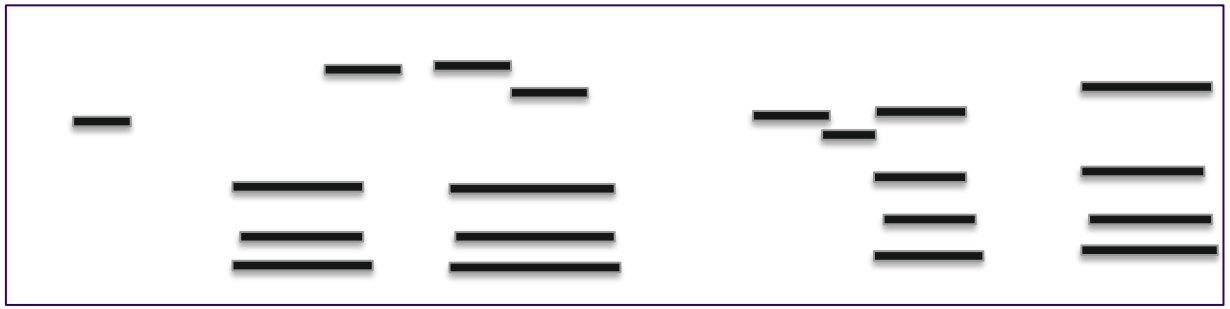


Prior work: Cascades

Score



Performance MIDI



Audio



PM2S:

Performance MIDI to Score "Transcription"

MIDI

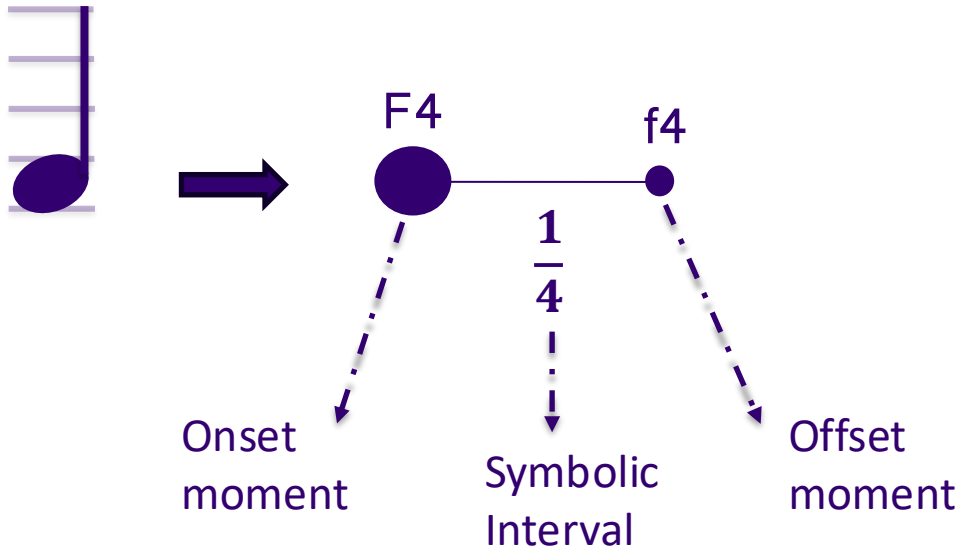
"Transcription"

Full transcription

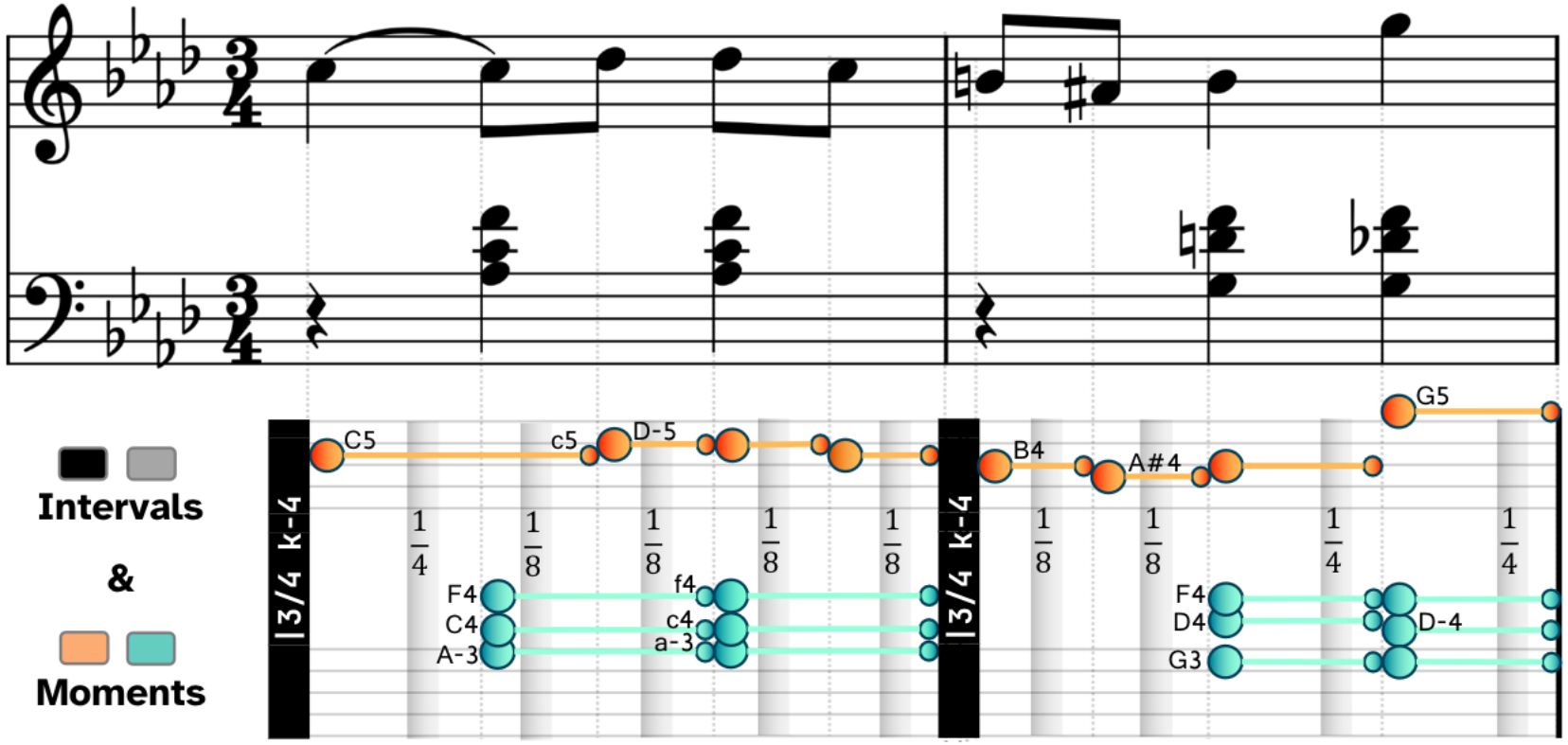
Intervals and Moments (InterMo

Symbolic intervals: Barlines and metric intervals

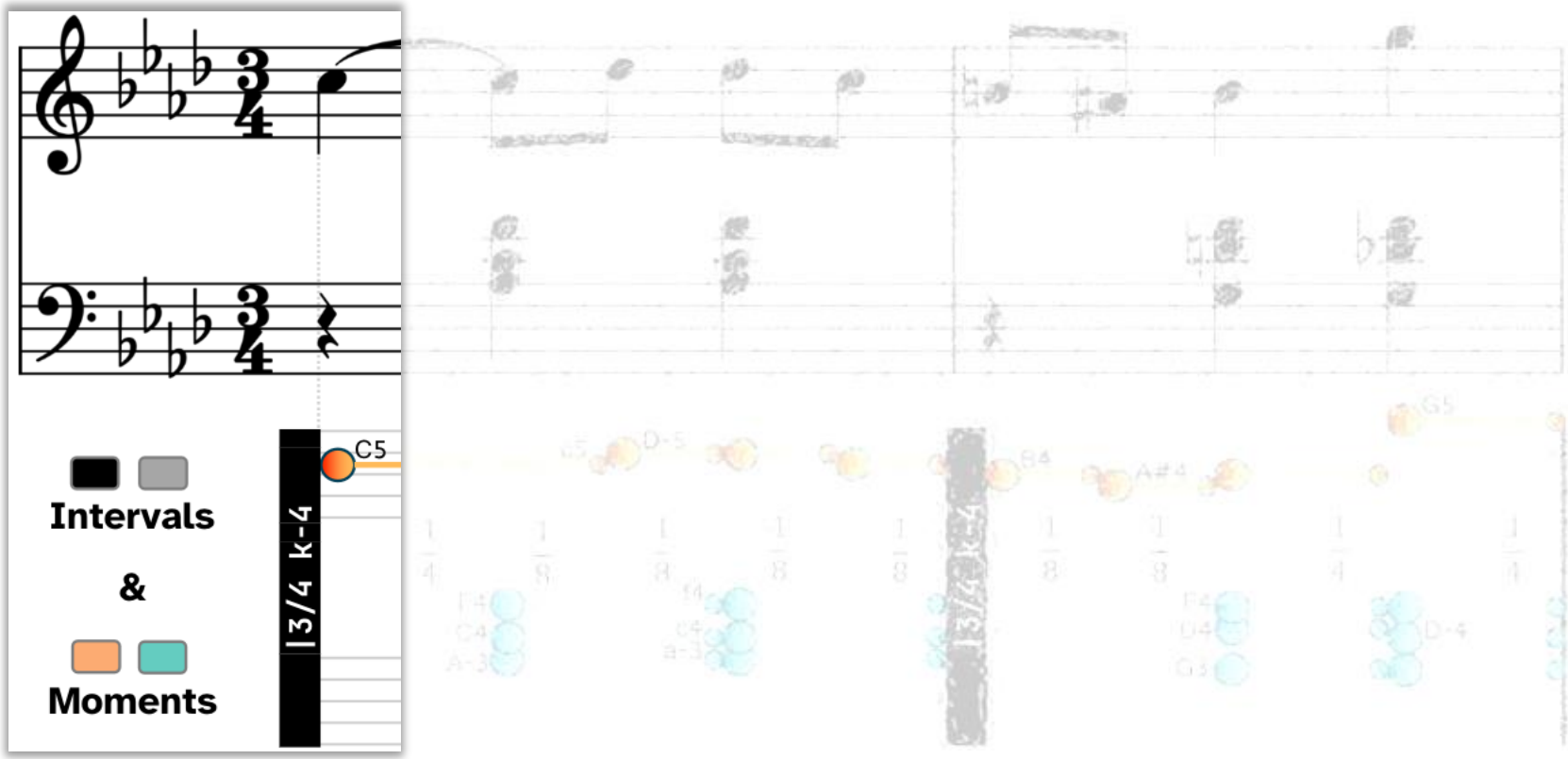
Momentary state changes: Onsets and offsets



Intervals and Moments (InterMo



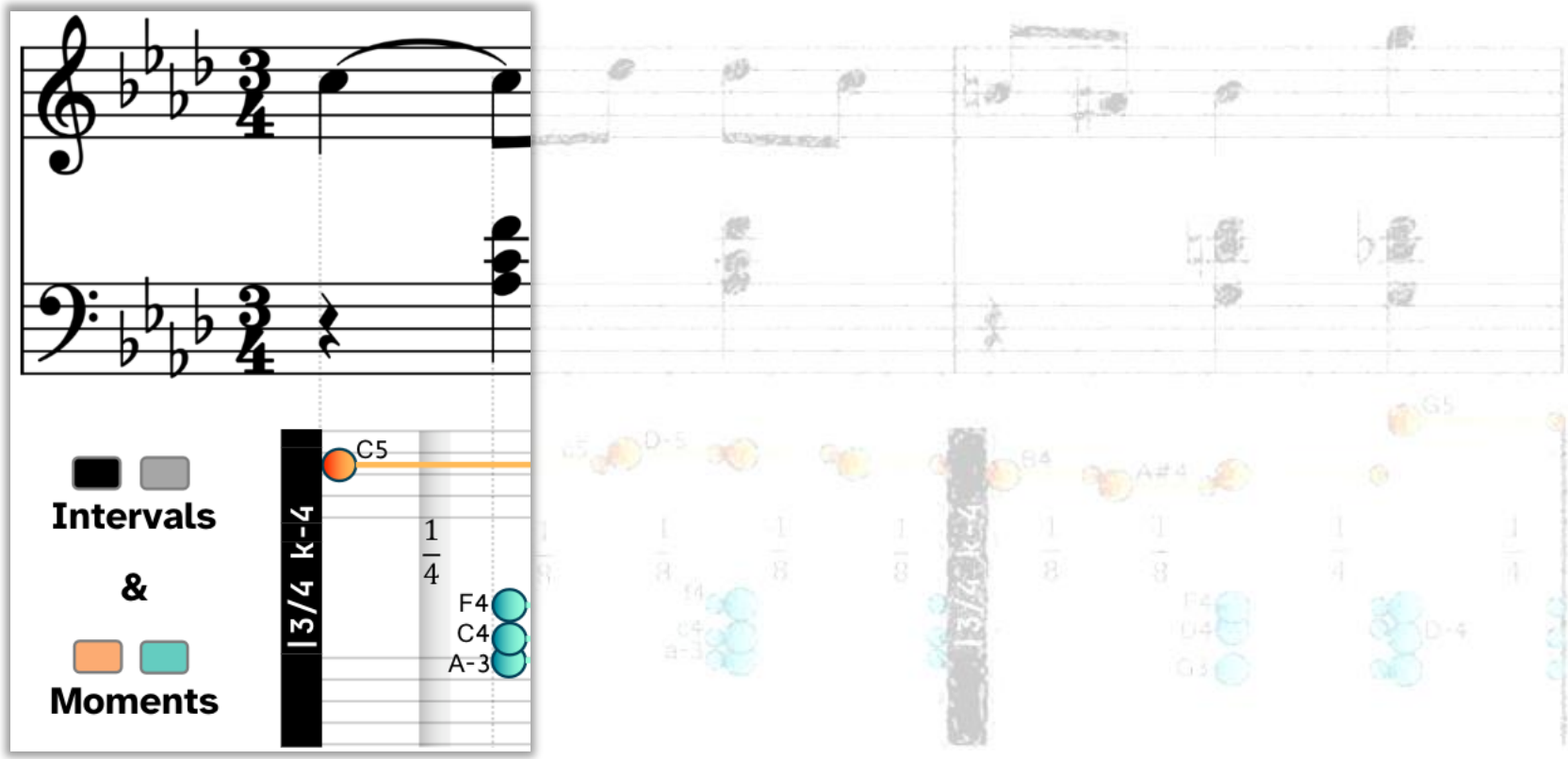
InterMo unites notation with grounding.



|3/4k-4 New bar, 3/4 time, 4 flats. PR: (piano-right) press C5.

at t=0.2s

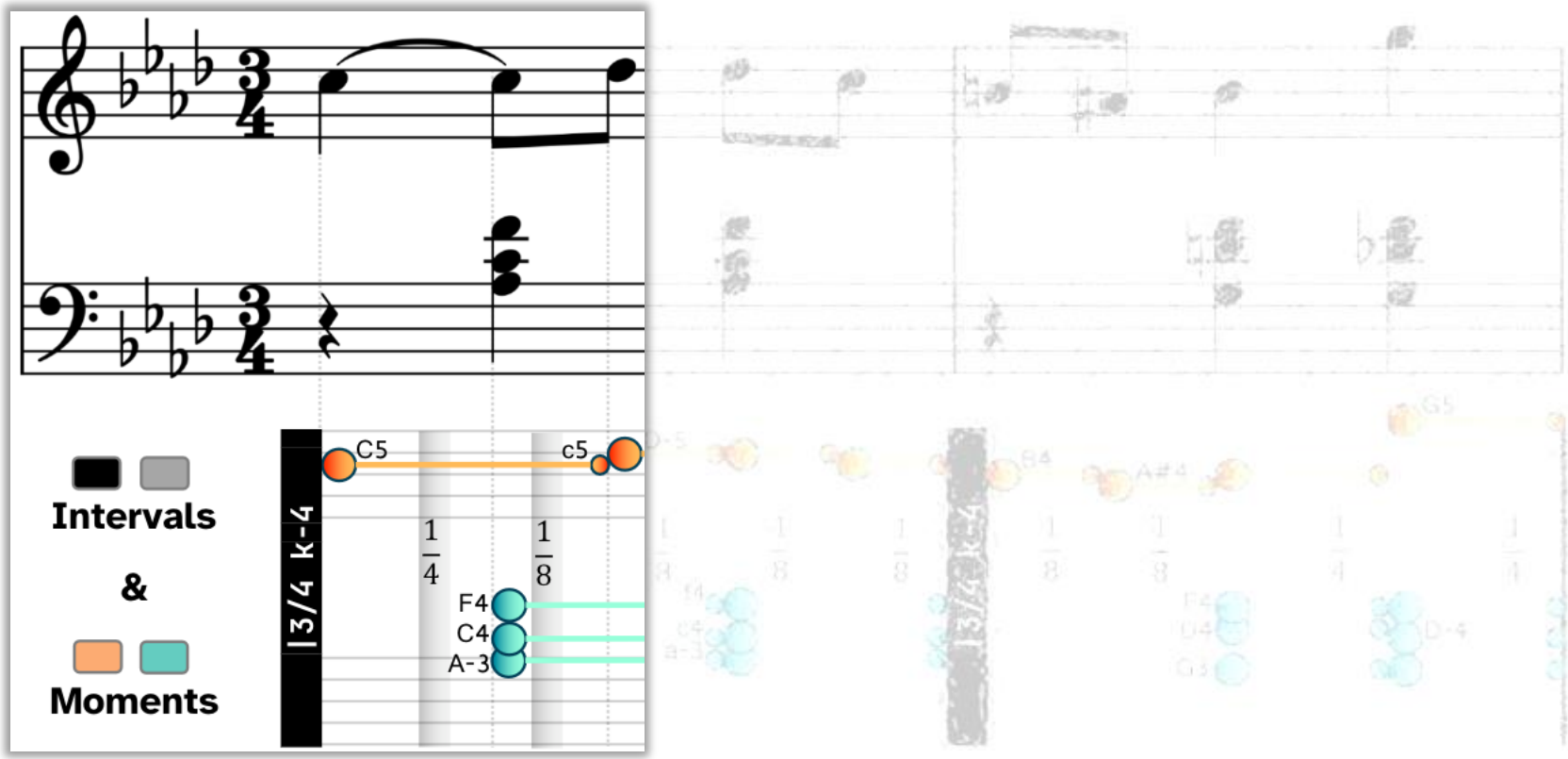
InterMo unites notation with grounding.



1/4 Wait a quarter note. PL: Press A-3 C4 F4 (sorted by pitch).

at t=1.4s

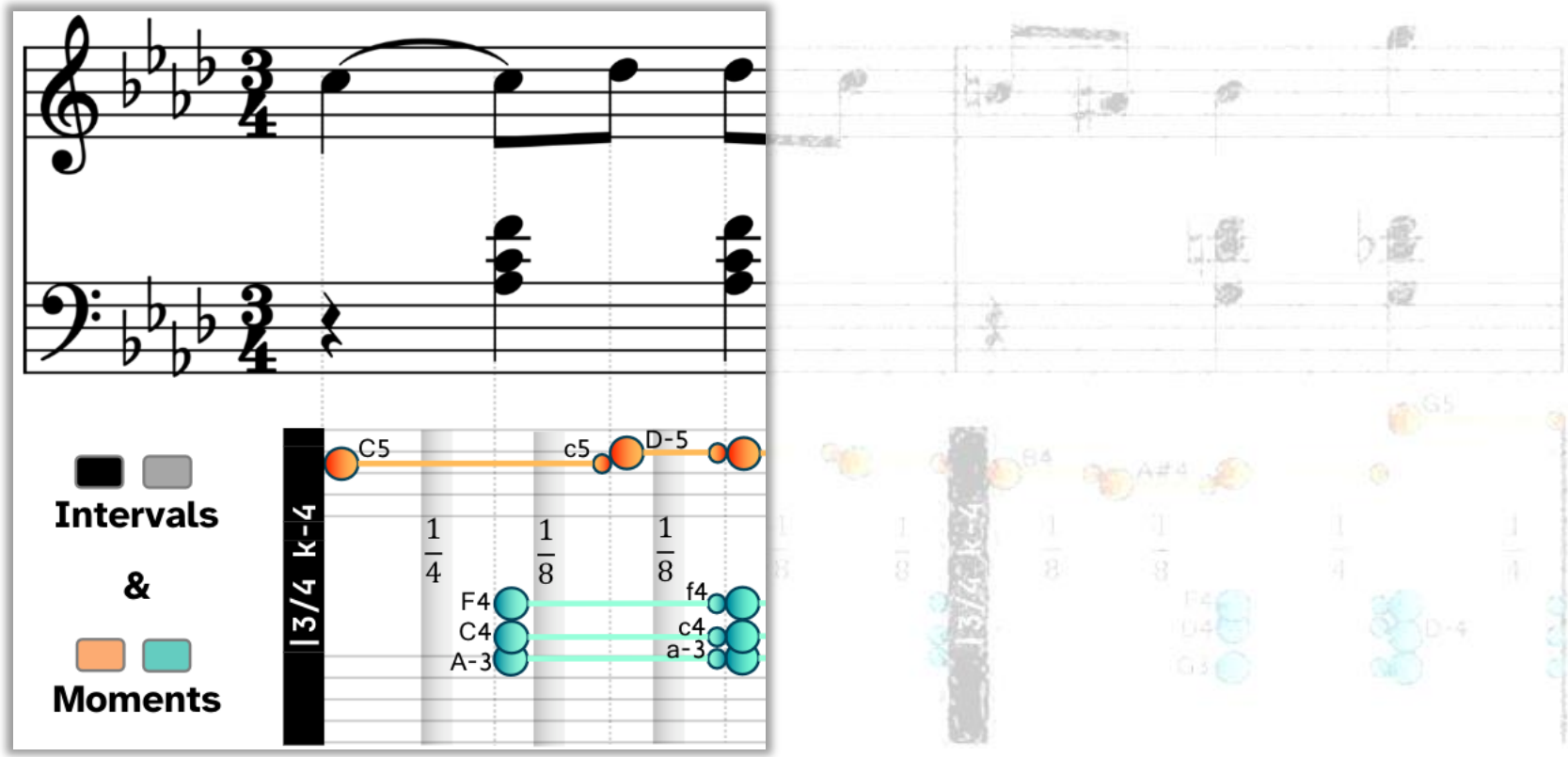
InterMo unites notation with grounding.



1/8 Wait an eighth note. PR: Release c5, press D-5.

at t=1.8s

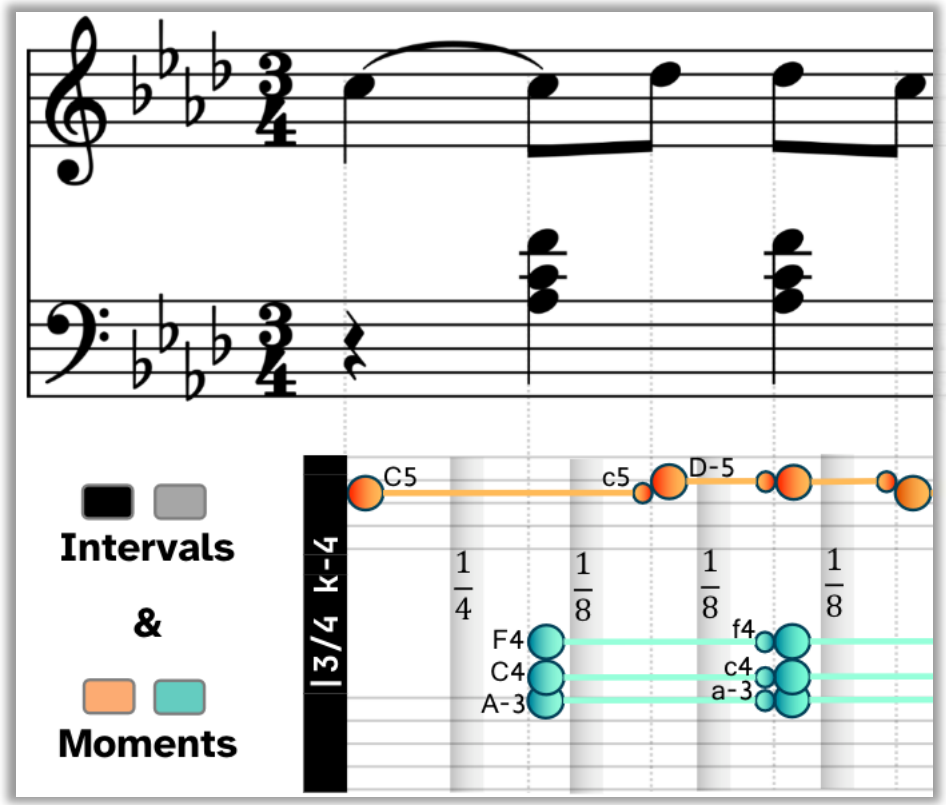
InterMo unites notation with grounding.



1/8 PL: Release a-3 c4 f4, re-press A-3 C4 F4. PR: Release d-5, re-press D-5. (staff → offsets before onsets → pitch)

at t=2.1s

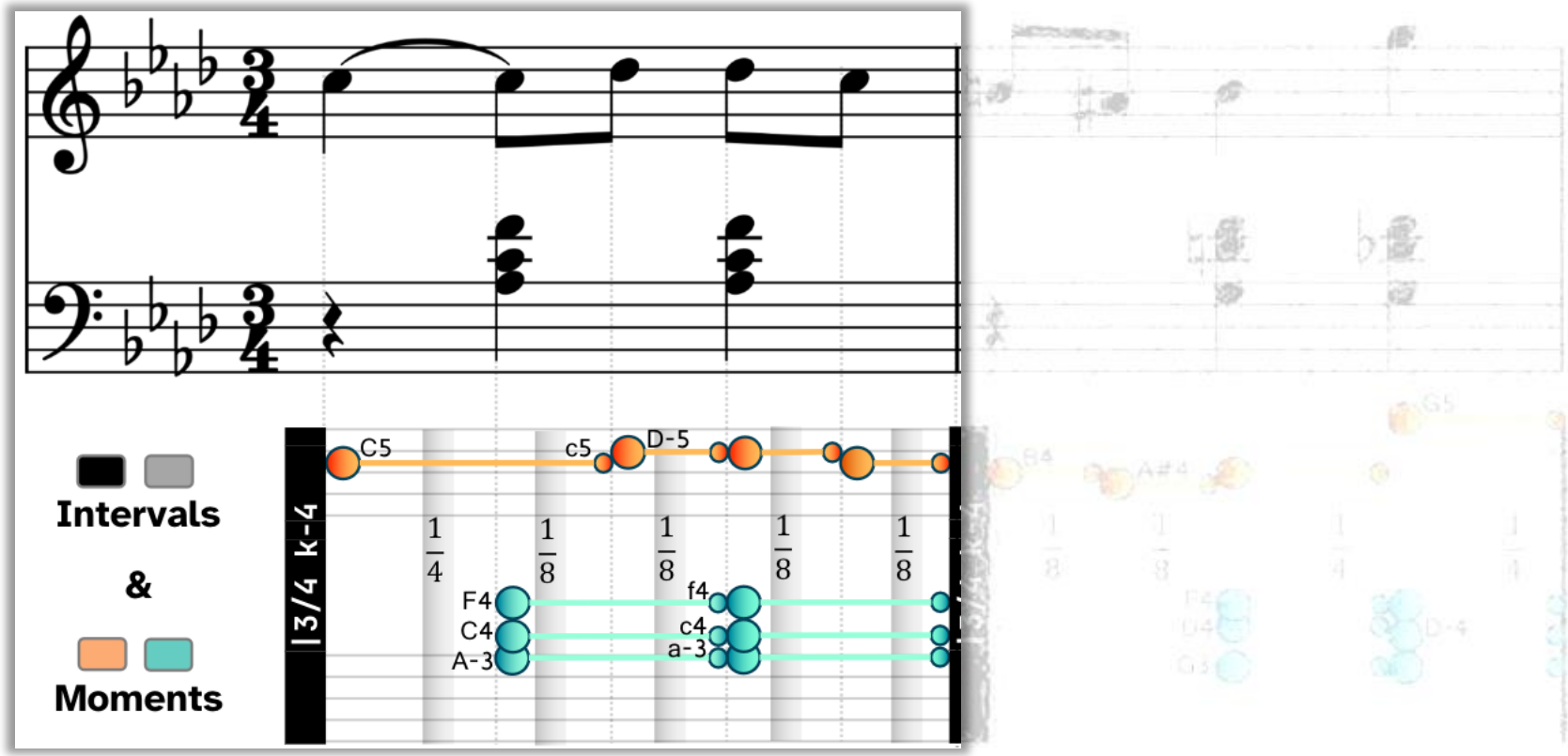
InterMo unites notation with grounding.



1/8 wait. Release d-5, press C5 (PR: still active).

at t=2.7s

InterMo unites notation with grounding.

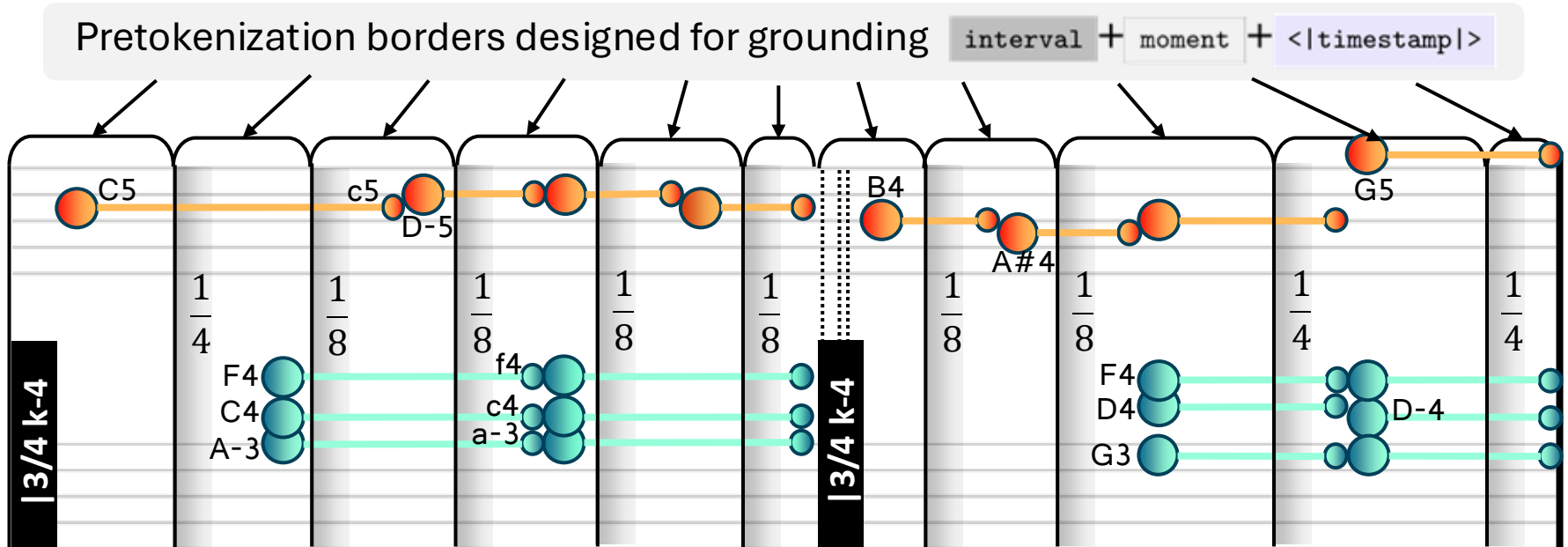


1/8 wait. On PL: Release a-3 c4 f4. On PR: release c5.

at t=3.2s

Intervals and Moments are tokenized together.

InterMo: |3/4k-4 PR: C5 1/4 PL: A-3 C4 F4 1/8 PR: c5 D-5 1/8 PL: a-3 c4 f4 A-3 C4 F4 PR: d-5 D-5 1/8 d-5 C5 1/8 PL: a-3 c4 f4 PR: c5 |3/4k-4 B4 1/8 b4 A#4 1/8 PL: G3 D4 F4 PR: a#4 B4 1/4 PL: g3 d4 f4 G3 D-4 F4 PR: b4 G5 1/4 PL: g3 d-4 f4 PR: g5



Different configurations of Intervals and Moments correspond to different tasks

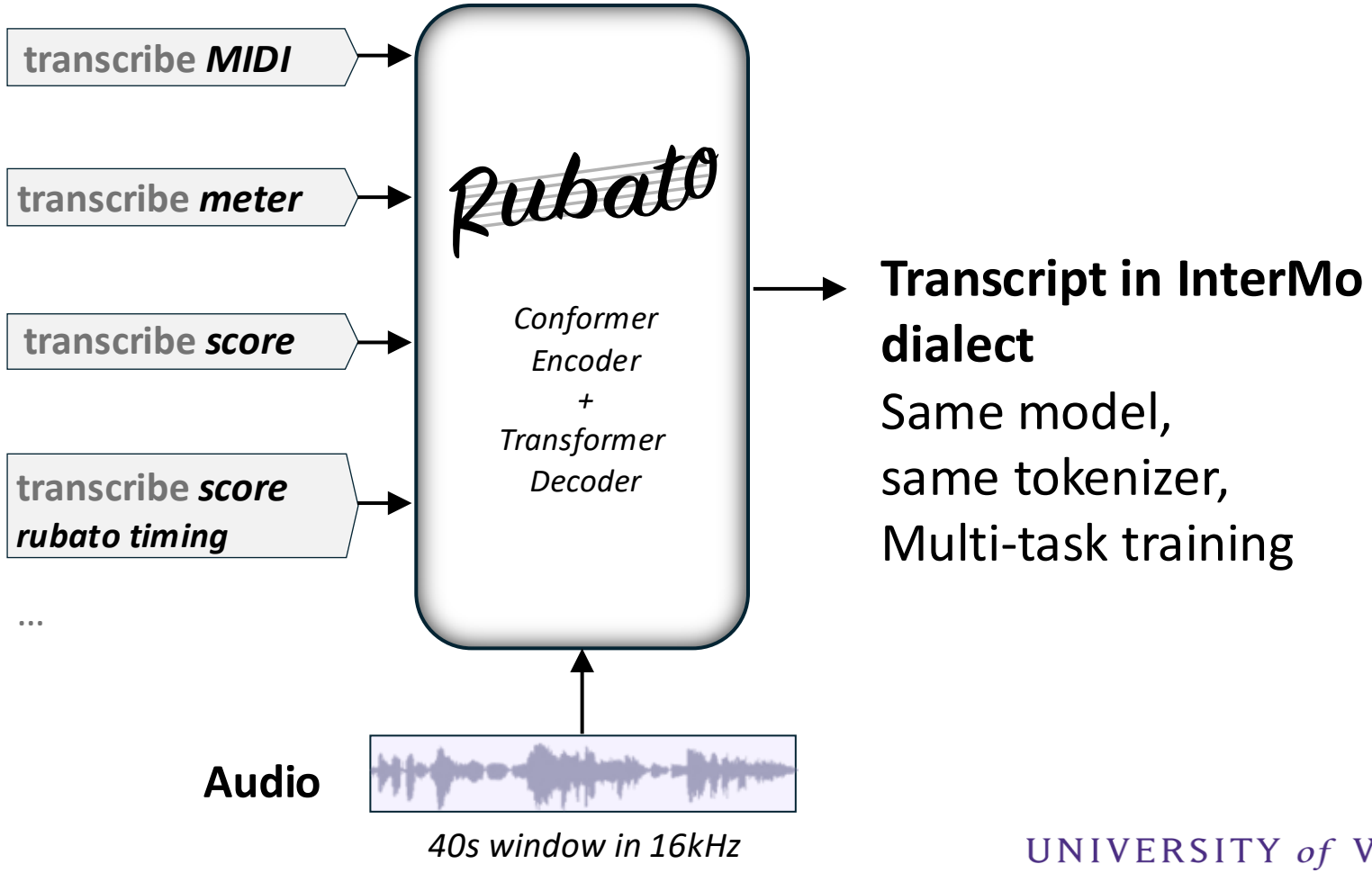
Metrical positions can be analyzed separately.



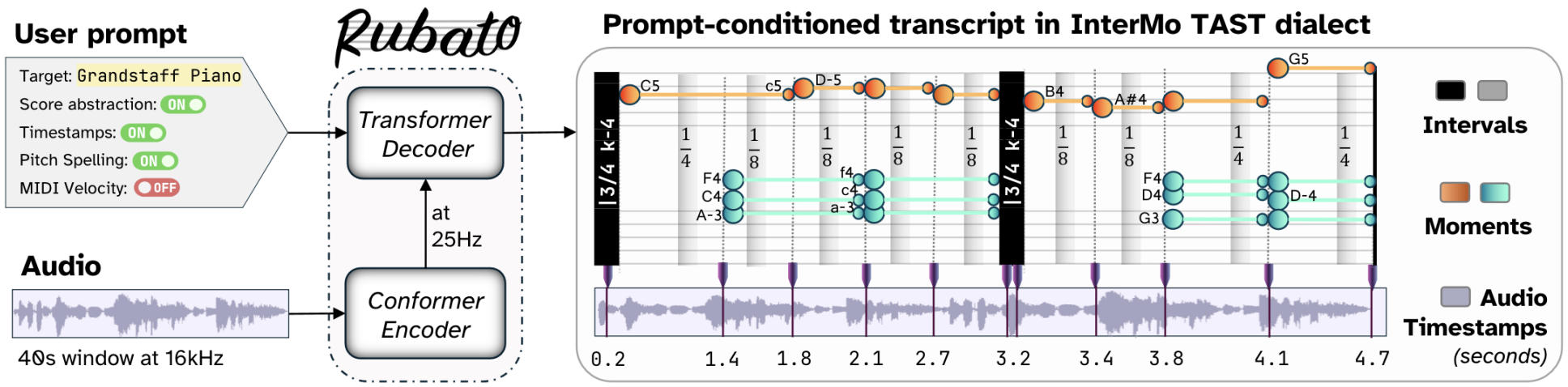
Note state tracking is separate from symbolic time.

Timestamps follow monotonic order and encode expressive timing.

Rubato: Prompt Conditioned Transcription



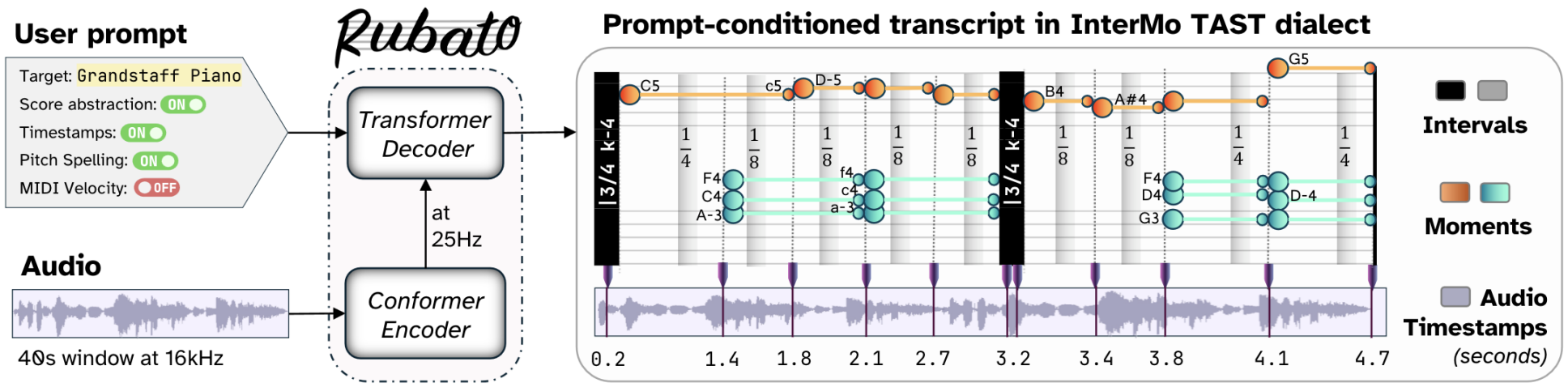
Prompt-conditioned Transcription



Dialect	Prompt → Target example (first bar of figure above)
TAST	`<|piano|> <|score|> <|timestamp|> <|spell|>` → `|3/4k-4 PR: C5 <|0.20|> 1/4 PL: A-3 C4 F4 <|1.40|> 1/8 PR: c5 D-5 <|1.80|>` `1/8 PL: a-3 c4 f4 A-3 C4 F4 PR: d-5 D-5 <|2.10|> 1/8 d-5 C5 <|2.70|>` `1/8 PL: A-3 C4 F4 PR: c5 <|3.20|> |3/4k-4 B4 <|3.20|> ...`
TAST$_{lite}$	`<|piano|> <|score|> <|timestamp|>` → `|3/4k-4 PR: C5 <|0.20|> 1/4 PL: G#3 C4 F4 <|1.40|> 1/8 PR: c5 C#5 <|1.80|>` `1/8 PL: g#3 c4 f4 G#3 C4 F4 PR: c#5 C#5 <|2.10|> 1/8 c#5 C5 <|2.70|>` `1/8 PL: G#3 C4 F4 PR: c5 <|3.20|> |3/4k-4 B4 <|3.20|> ...`

***T**ime **A**ligned **S**core **T**ranscription: full score notation aligned to audio through timestamps (TAST); without notational pitch spelling (TAST$_{lite}$).*

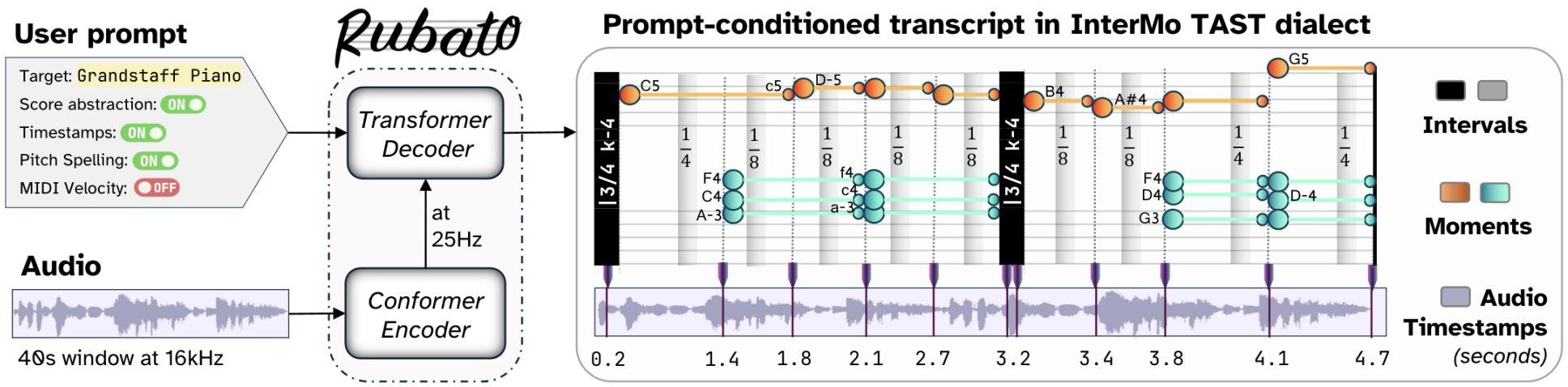
Prompt-conditioned Transcription



Dialect	Prompt → Target example (first bar of figure above)
A2S	`<|piano|> <|score|> <|spell|>` → `|3/4k-4 PR: C5 1/4 PL: A-3 C4 F4 1/8 PR: c5 D-5 1/8 PL: a-3 c4 f4 A-3 C4 F4 PR: d-5 D-5 1/8 d-5 C5 1/8 PL: a-3 c4 f4 PR: c5 |3/4k-4 B4 ...`
A2S$_{lite}$	`<|piano|> <|score|>` → `|3/4k-4 PR: C5 1/4 PL: G#3 C4 F4 1/8 PR: c5 C#5 1/8 PL: g#3 c4 f4 G#3 C4 F4 PR: c#5 C#5 1/8 c#5 C5 1/8 PL: g#3 c4 f4 PR: c5 |3/4k-4 B4 ...`

Audio-to-Score Transcription: score notation matching §2 (A2S), same without pitch spelling (A2S$_{lite}$)—Training only.

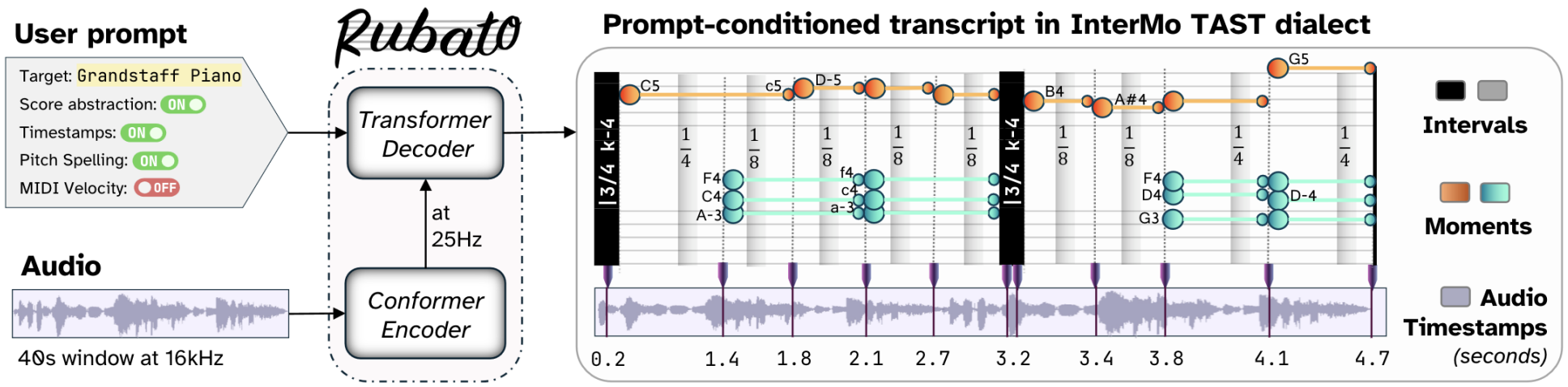
Prompt-conditioned Transcription



Dialect	Prompt → Target example (first bar of figure above)
AMT	`<|piano|> <|timestamp|> <|MIDI|>` → `<|CC64:on|> <|0.10|> C5 <|vel:46|> <|0.20|> G#3 <|vel:49|> <|1.37|> F4 <|vel:43|> <|1.40|> C4 <|vel:56|> <|1.41|> c5 <|1.61|> <|CC64:off|> <|1.70|> C#5 <|vel:46|> <|1.80|> g#3 f4 <|2.00|> c#5 <|2.01|> c4 <|2.02|> G#3 <|vel:36|> <|2.08|> C#5 <|vel:44|> <|2.10|> F4 <|vel:46|> C4 <|vel:39|> <|2.11|> C5 <|vel:43|> <|2.70|> c#5 <|2.72|> c5 <|3.16|> g#3 f4 <|3.17|> c4 <|3.18|> ...`

Automatic MIDI Transcription: pitched events with timestamp grounding (AMT_{lite}); with MIDI-specific "velocity" and pedal controls (AMT)

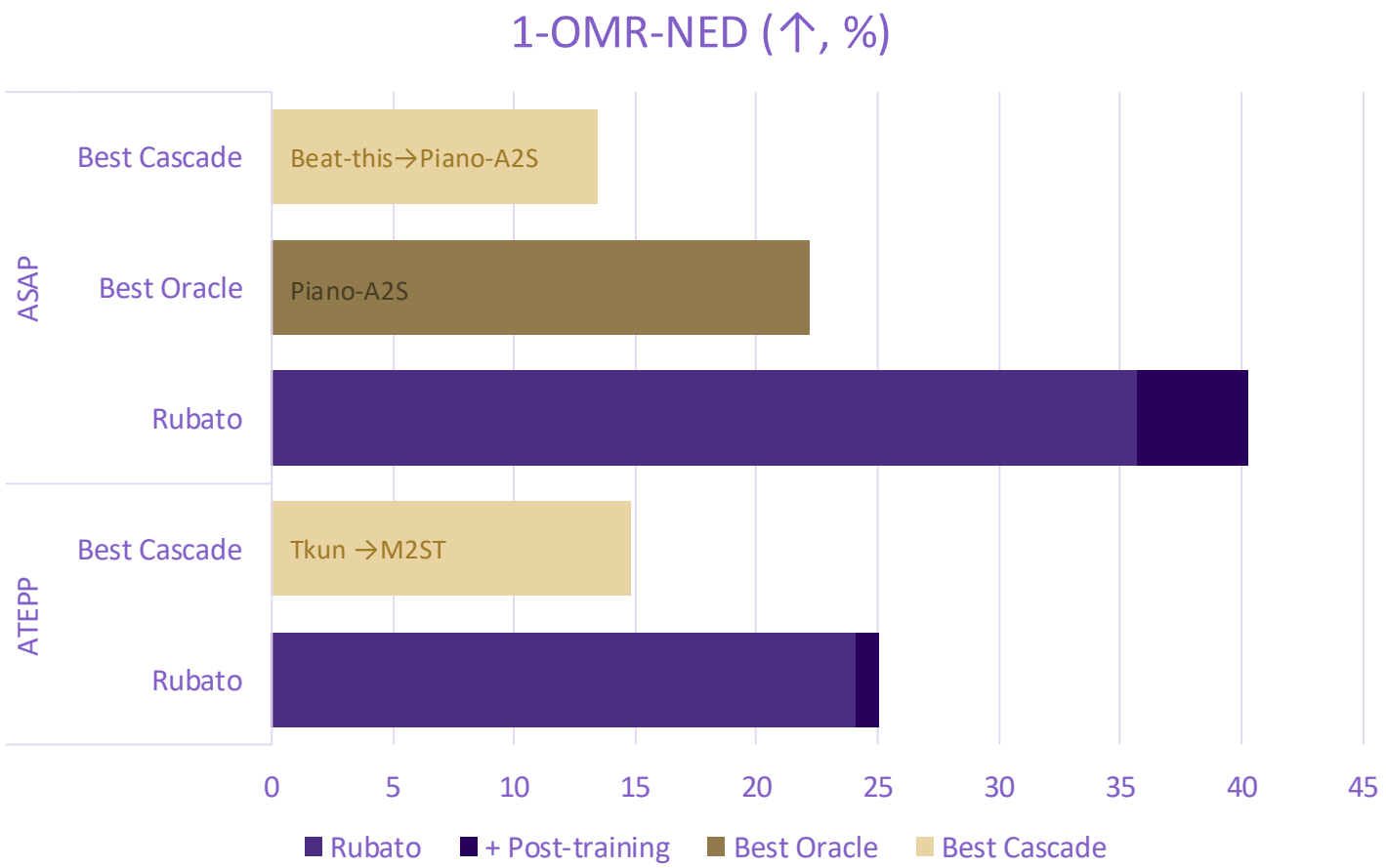
Prompt-conditioned Transcription



Dialect	Prompt → Target example (first bar of figure above)
DBD_{plus}	<|beats|> <|score|> <|timestamp|> → |3/4k-4 <|0.20|> 1/4 * <|1.40|> 1/4 * <|2.10|> 1/4 |3/4k-4 <|3.20|> ...
DBD	<|beats|> <|timestamp|> → | <|0.20|> * <|1.40|> * <|2.10|> | <|3.20|> ...

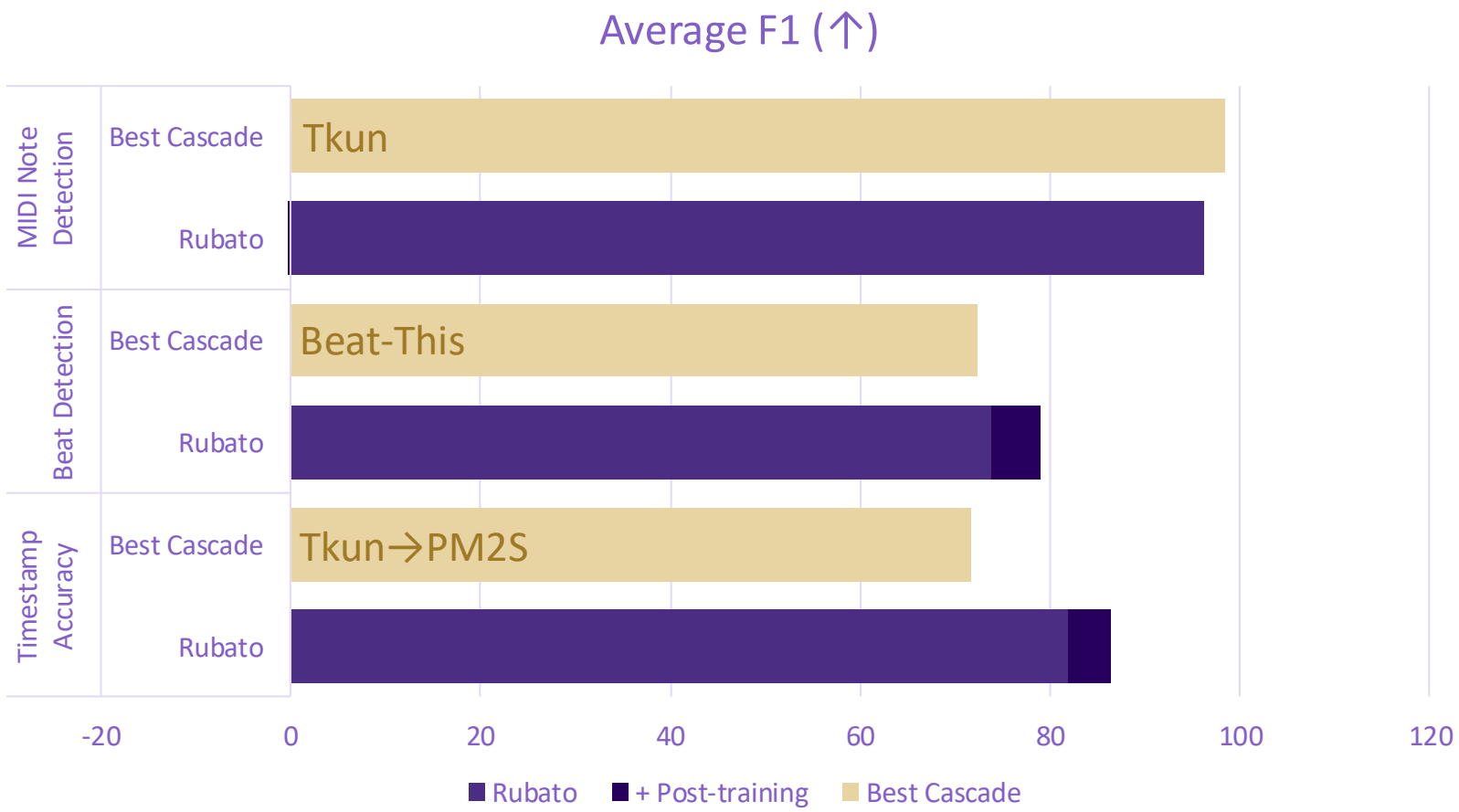
***D**ownbeat and **B**eat **D**etection: rhythmic anchors in the metrical scaffold (DBD_{plus}); beat (*) and downbeats (|) without meter and key (DBD).*

Experiments – Sheet Music Quality



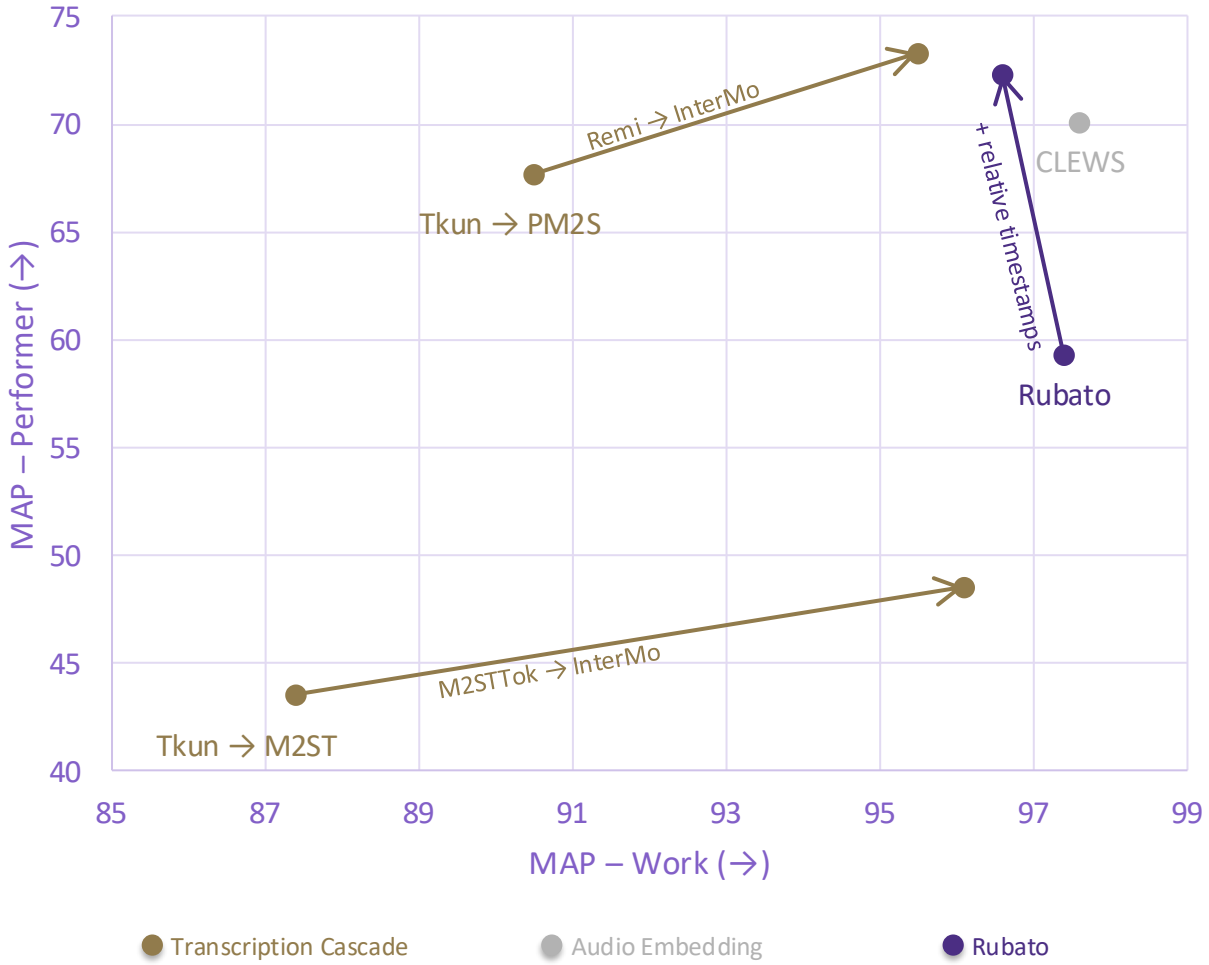
- Even when cascaded systems are given ground truth MIDI, Rubato produces higher quality sheet music
- Further post-training widens the gap

Experiments – Timestamp Grounding Accuracy



- Our single model performs on par with specialists at timestamp based tasks

Work and Performer Identity Preservation



- Rubato outperforms all cascades in their native tokenization format, and is comparable with audio embedding based version identification methods
- Adding relative timestamps increases the performer identification ability while only minimally decreasing the work identification ability
- When cascades are retokenized with InterMo, work and performer identity are better preserved

Rubato + InterMo

- Higher quality sheet music
- Notation-native timing information
- Grounded sheet music generation

Read our paper!

Questions?

Contact: nctamer@cs.washington.edu

Explore our demo!