High-Resolution Violin Transcription using Weak Labels

We present the *Multi-Stream Conformer (MUSC)*, a SOTA violin transcriber that converts **44.1 kHz raw audio** into **MIDI with 5.8ms time- and 10-cent frequency-resolution**, and without requiring frame-wise labels during training!

SCAN FOR GITHUB & DATASET

upf. MTG Music Technology Group

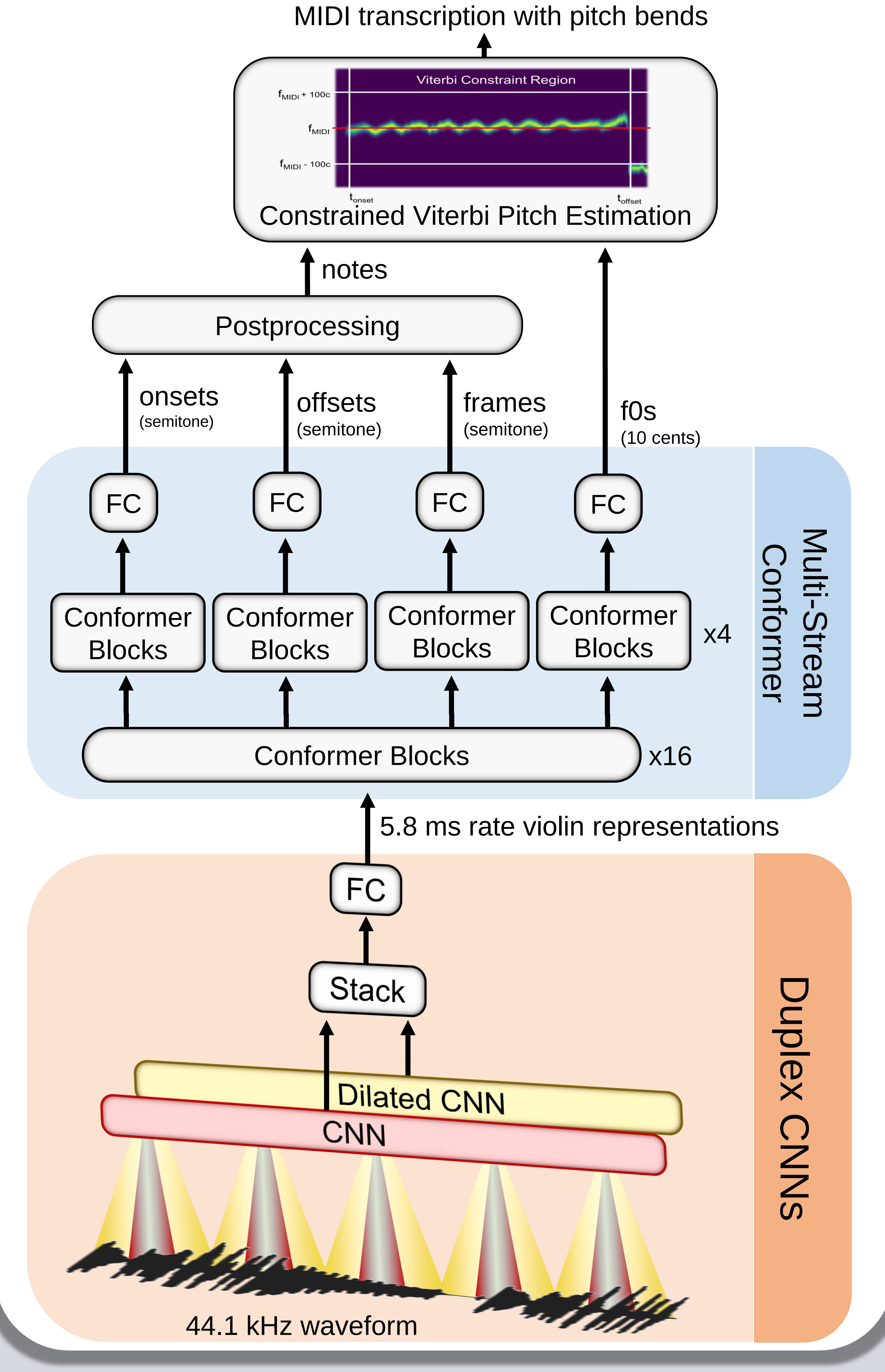
Nazif C. Tamer[♭], **Yigitcan Özer**[♯], **Meinard Müller**[♯], **Xavier Serra**[♭]

♭ Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain, ♯ International Audio Laboratories Erlangen, Germany

AUDIO LABS

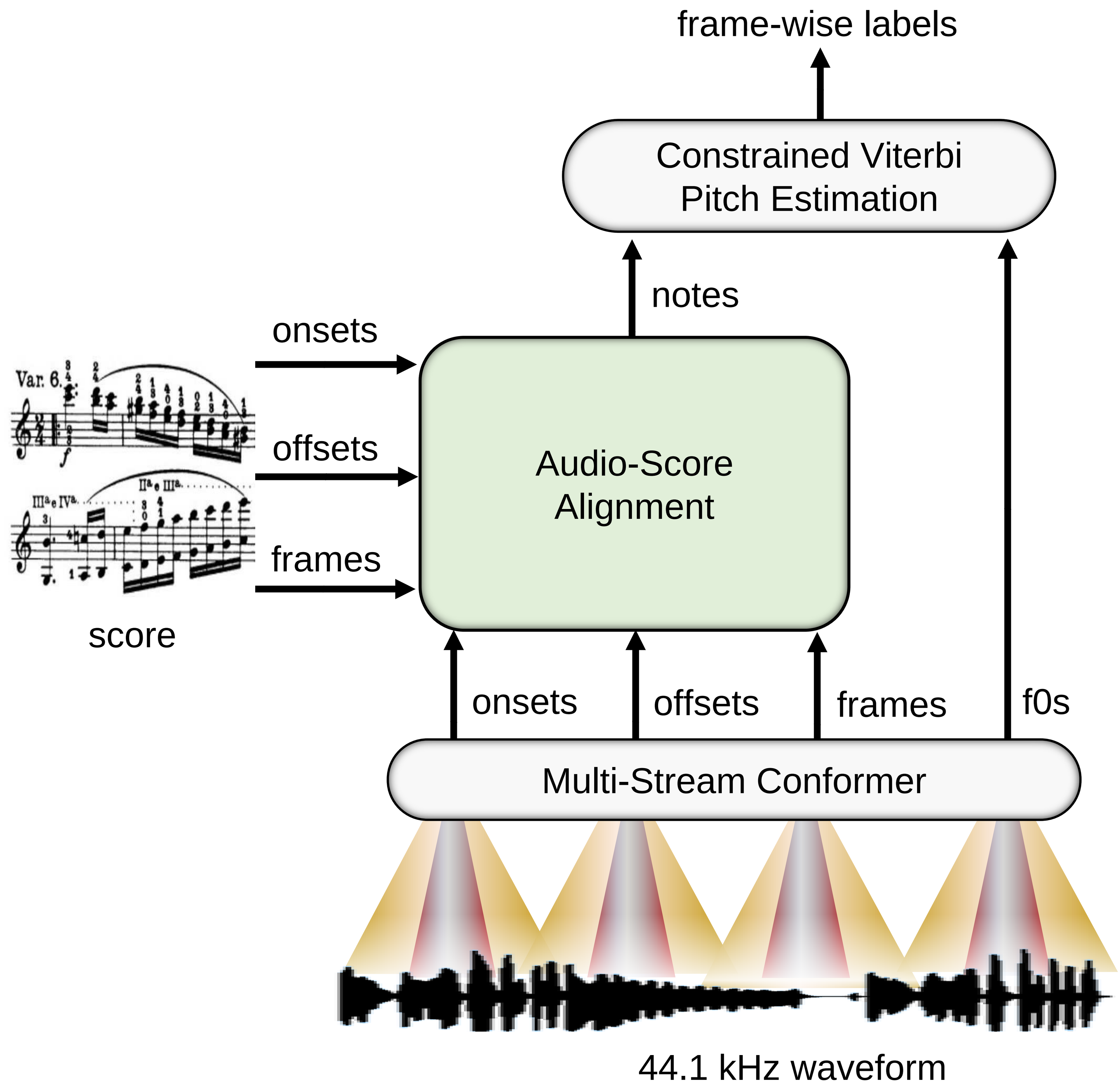
Architecture

CNNs and Conformer blocks convert raw audio into 4 representations, then postprocessing creates the MIDI with pitch bends.



Training

As a large-scale violin dataset with frame-wise labels do not exist, MUSC generates its frame-wise training labels by aligning its own onset, offset, and frame feature representations with music score.



Dataset: It is trained on 120 violin etudes from three books & their unaligned YouTube recordings.

Method	Etudes	Players	Performances	Duration (h)
Paganini, Op. 1	24	10	235	13:00
Wohlfahrt, Op. 45	60	6	506	11:36
Kayser, Op. 20	41	8	280	09:48
Total	**120**	**22**	**1021**	**34:23**

The released dataset can be found on GitHub, with the frame-wise alignments generated by our model.

Tests yield SOTA performance for two proxy tasks: Violin Transcription & Pitch Estimation

Transcription

Compared with MT3[1] and Basic Pitch[2] on two datasets. (URMP is involved in MT3 training set.)

	URMP				Bach10			
	P	R	F1	$F1_{no}$	P	R	F1	$F1_{no}$
MUSC	**86.5**	83.1	**84.6**	**93.0**	**65.0**	**64.8**	**64.8**	**77.0**
MT3[1]	79.1	**87.1**	82.2	88.9	54.2	51.5	52.7	62.0
BP[2]	58.8	67.9	62.8	83.3	33.6	43.2	37.6	57.5

[1] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. H. Engel, "MT3: multi-task multitrack music transcription," in CoRR, 2021.
[2] R. M. Bittner et al., "A lightweight instrument agnostic model for polyphonic note transcription and multipitch estimation," in Proc. ICASSP, 2022.
[3] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in Proc. ICASSP, 2018.
[4] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music." in JASA, 2002.
[5] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in Proc. ICASSP), 2014.
[6] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," in JASA, 2008.

Pitch Estimation

Compared with CREPE[3], YIN[4], pYIN[5], and SWIPE[6] (v: Viterbi post-processing, Bach10 is involved in CREPE training set)

	URMP		Bach10	
	RPA50	RPA10	RPA50	RPA10
MUSC	98.3	89.0	98.3	86.9
vMUSC	**98.6**	**89.4**	98.4	87.0
CREPE[3]	96.4	87.2	**98.6**	**88.1**
vCREPE	97.3	88.4	**98.6**	**88.1**
YIN[4]	95.3	88.4	97.1	81.7
pYIN[5]	97.2	88.6	97.4	80.3
SWIPE[6]	97.2	89.3	97.7	84.3