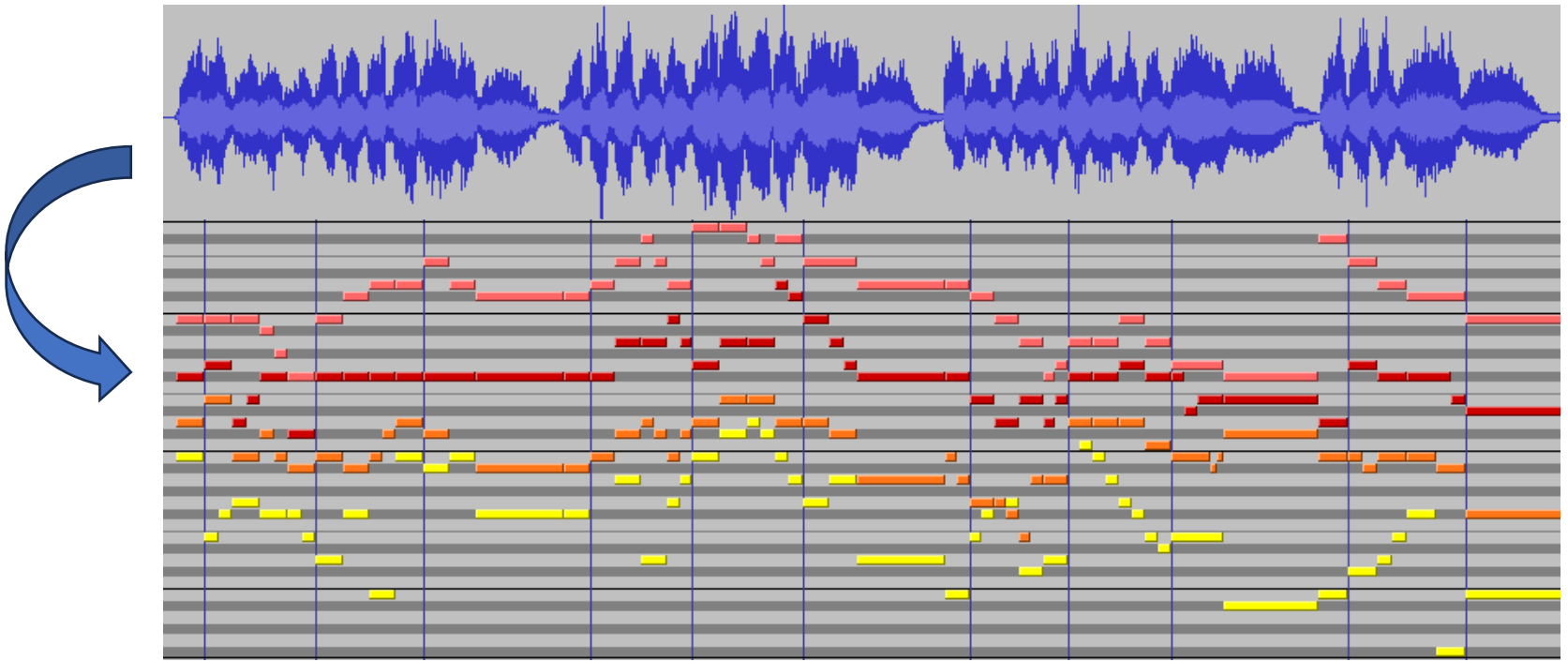
# Chorale Music Transcription with Soft-DTW Training Loss

Huiran Yu

Dec 13, 2023

# Problem Definition

- Given an audio, transcript the corresponding notes from each track



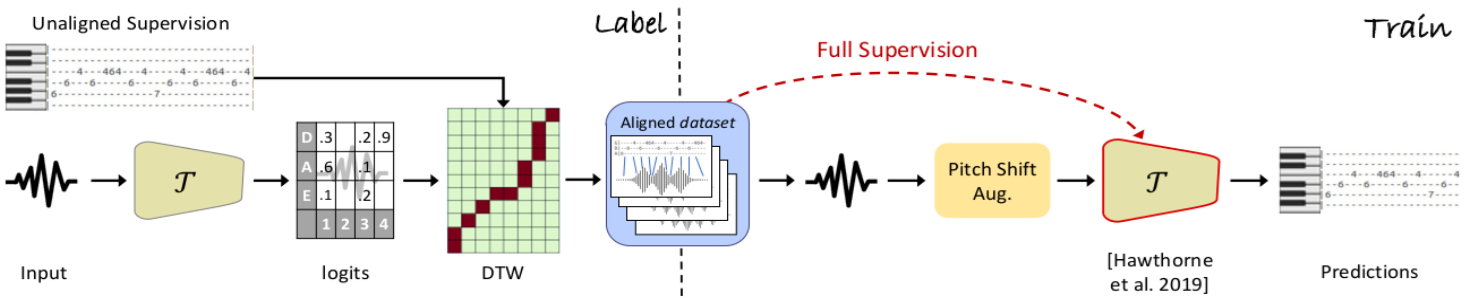
# What's Special about Chorale Music?

- **Lack of dataset**: Most of the datasets available are less than one hour, insufficient for large model training
- However, there exist recordings on YouTube and MIDI files in database which are **not timely aligned**

- To make use of these unaligned data, we need to find a way to do training with content-aligned, not timely-aligned supervision – **Dynamic Time Warping (DTW)**

# Previous Method

- B Maman et al. "Unaligned supervision for automatic music transcription in the wild." International Conference on Machine Learning. PMLR, 2022.



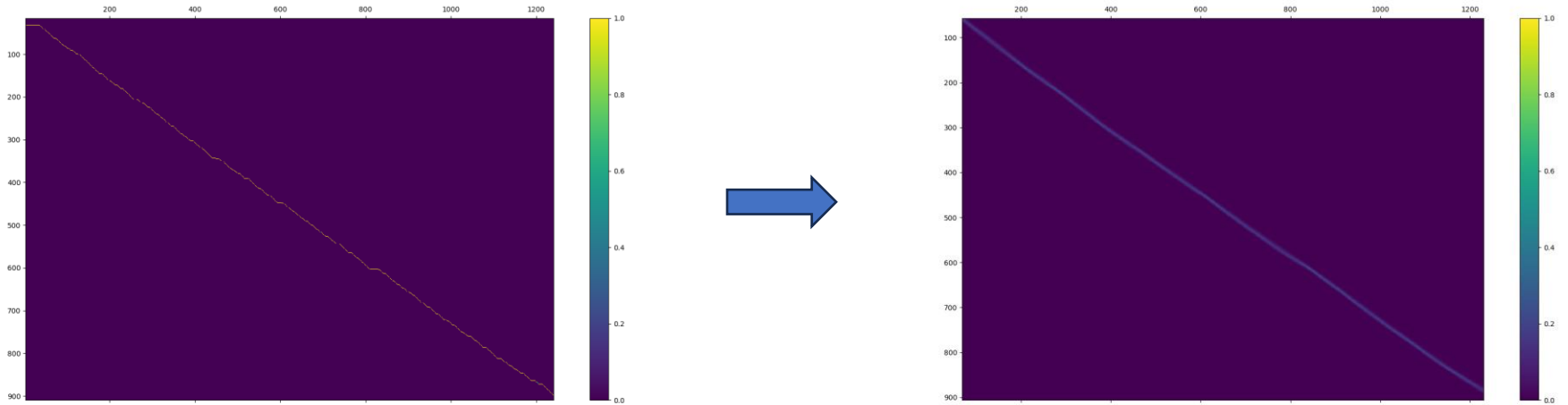

- **Pretrain** the model with **synthesized data**
- Perform **DTW between the transcription and the unaligned labels** to acquire training target

# Previous Method

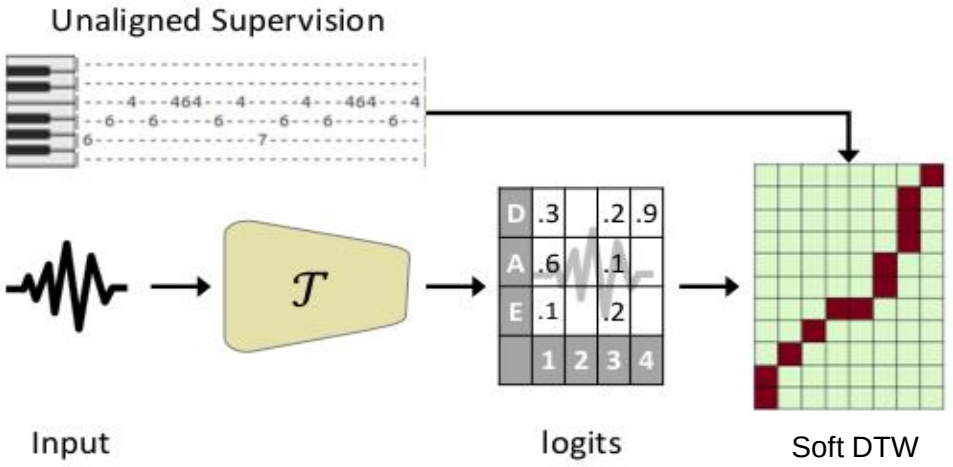
- Pro: Allowing the unaligned data to be used in training
- Con: **Not end-to-end**. The labeling process and the gradient descent are separated, since **DTW is not differentiable**.
- Is there any way to make gradient flow in DTW?

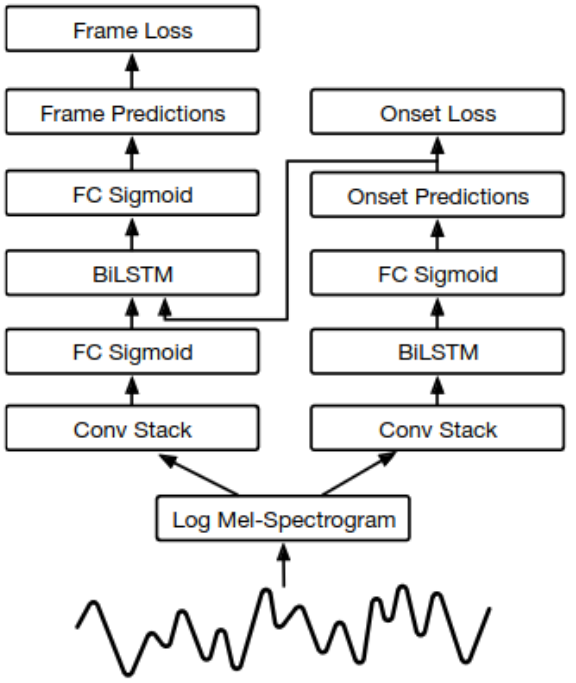
# Soft DTW

- Introduce gradient into the DTW alignment process
- Can be applied to measure the distance between prediction and unaligned ground-truth



# Architecture



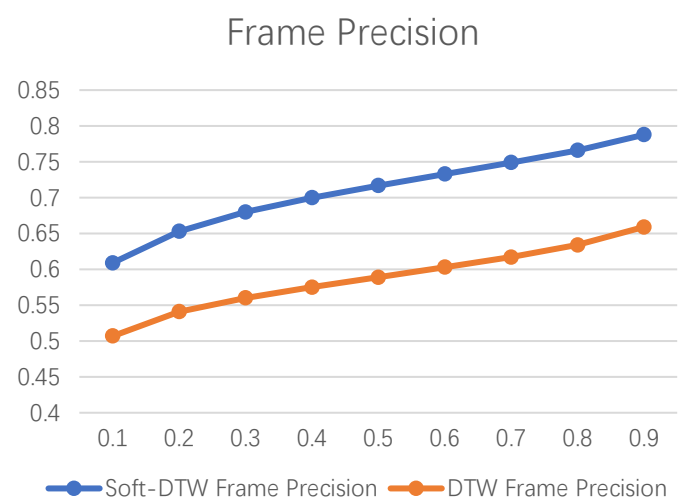
# Dataset and Representation

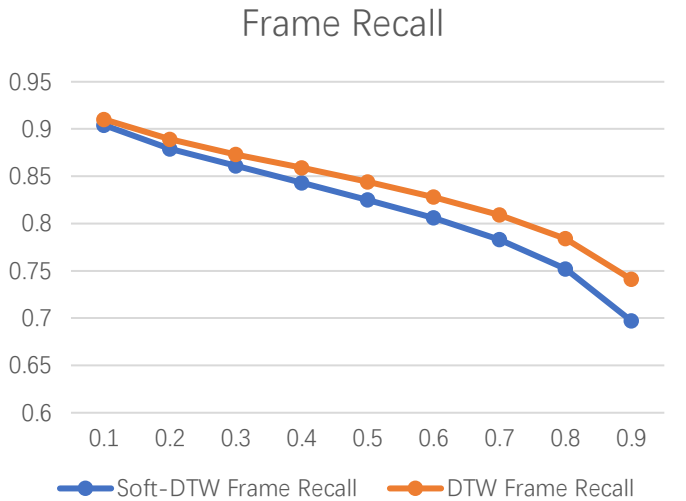
- **Bach Chorale**: 54 Bach composed chorale music recordings with corresponding midi, with total length of **1 hour 52 min**
- Using 47 of them; train:validation:test = 37:5:5
- Format
  - 3: onset; 2: sustain; 1: ending

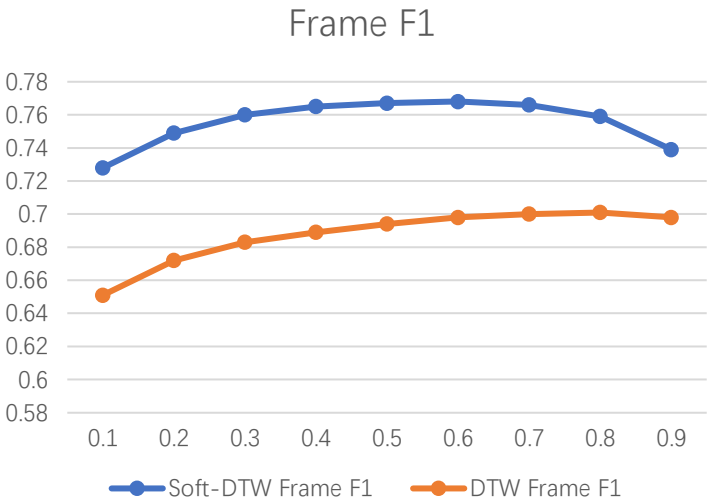
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 2 | 2 | 1 | | | | | | | |
| | | | | | | | | 3 | 1 | | |
| | | | | | 3 | 2 | 1 | | | | |
| | | | | | | | | | | 3 | 1 |

# Result

- Comparing between the DTW method and the proposed Soft-DTW method

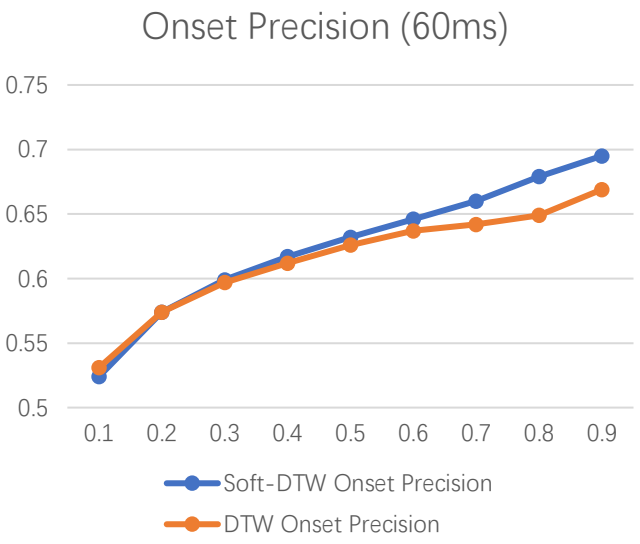


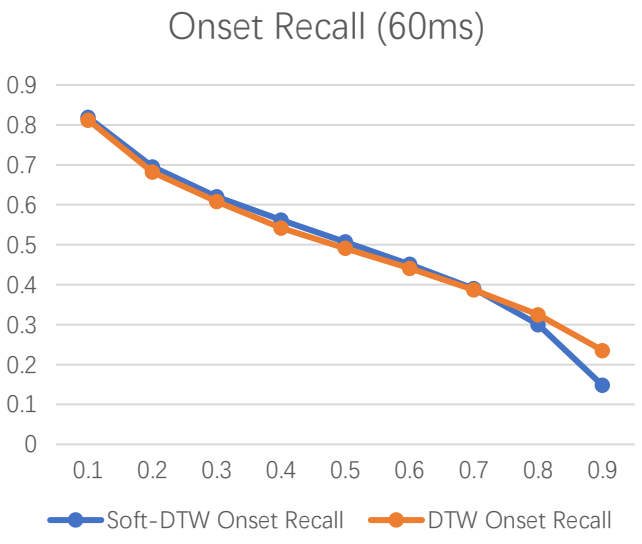



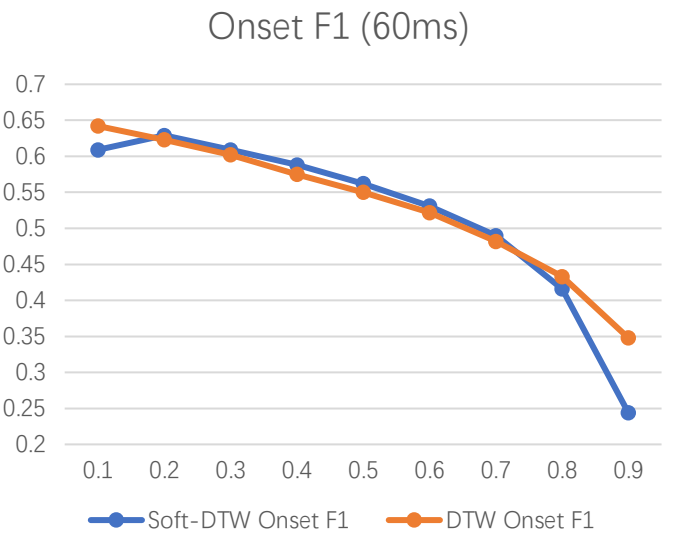




X axis is the threshold

# Result



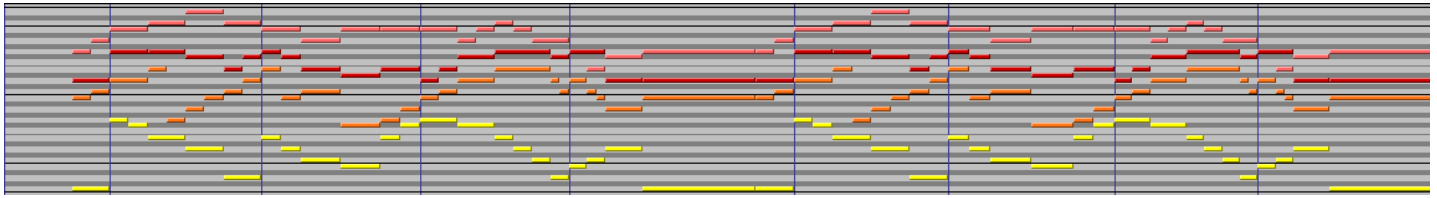



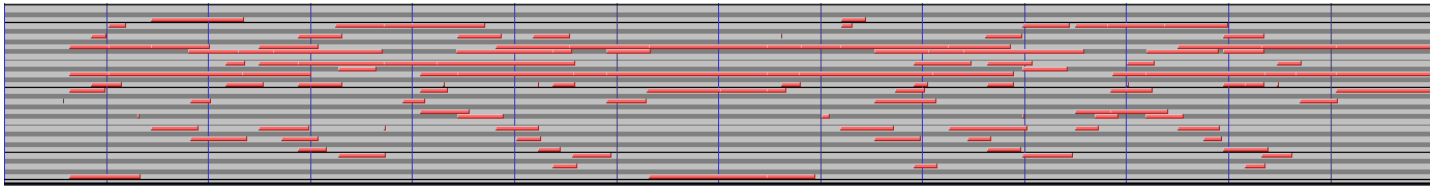

X axis is the threshold

# Result

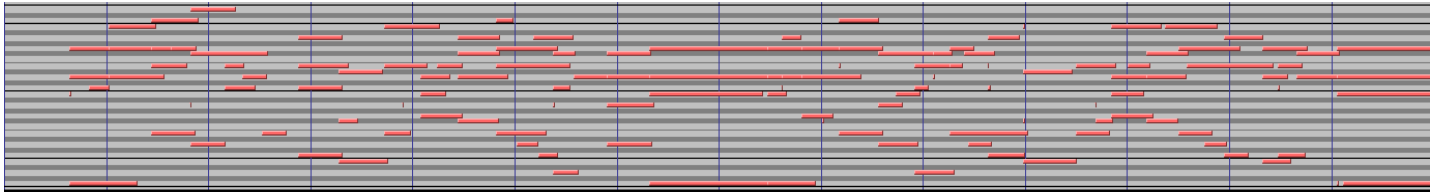
Input



DTW



Soft-DTW



# Future Works

- Collect larger dataset from the Internet to train the model
- Streaming: Identify different parts in performance; may involve a music content sequential model
- Generalize to general music transcription

# Thank you!