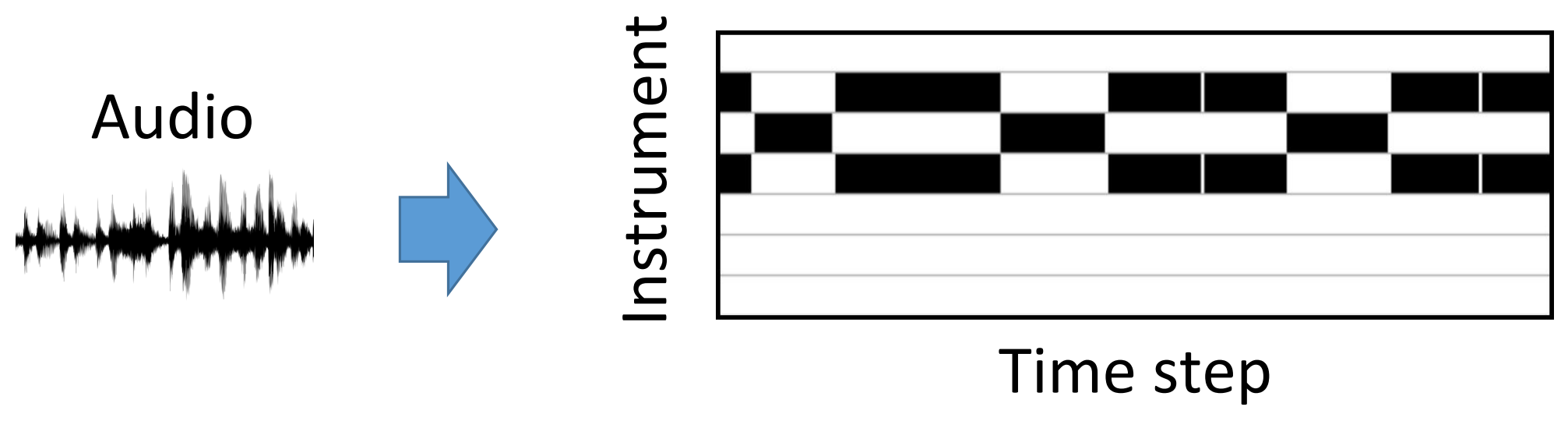
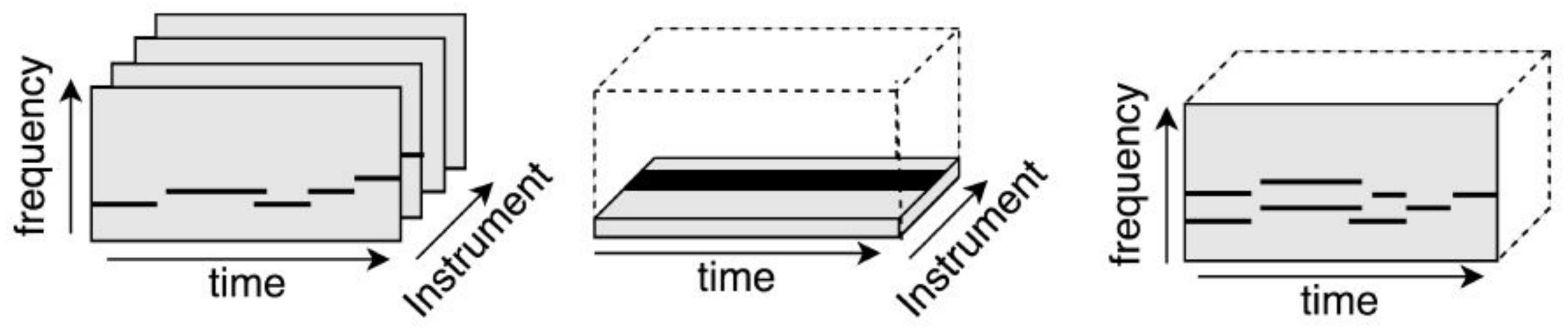
Multitask Learning For Frame-Level Instrument Recognition

[1] **Yun-Ning Hung**, [2] Yi-An Chen and [1] Yi-Hsuan Yang
[1] Research Center for IT innovation, Academia Sinica, Taipei, Taiwan
[2] KKBOX Inc., Taiwan

[Project Website] https://biboamy.github.io/streaming-demo/main_site/

Introduction





(a) Pianoroll (b) Instrument roll (c) Pitch roll

Frame-level instrument recognition

- Predict the instrument labels in each time frame
- Pitch can help frame-level instrument recognition [3]

Why multitask learning?

- By sharing representations between different tasks, we can enable our model to generalize better on our original task
- Has been used successfully across many applications, such as computer vision, NLP and speech recognition, but **not so much on music**

Multi-pitch streaming

- Predict the instrument that plays each individual note event (multi-pitch streaming)
- Piano roll: representation for multi-pitch streaming

Data

Problem

- No big dataset with instrument and pitch labels

Musescore dataset:

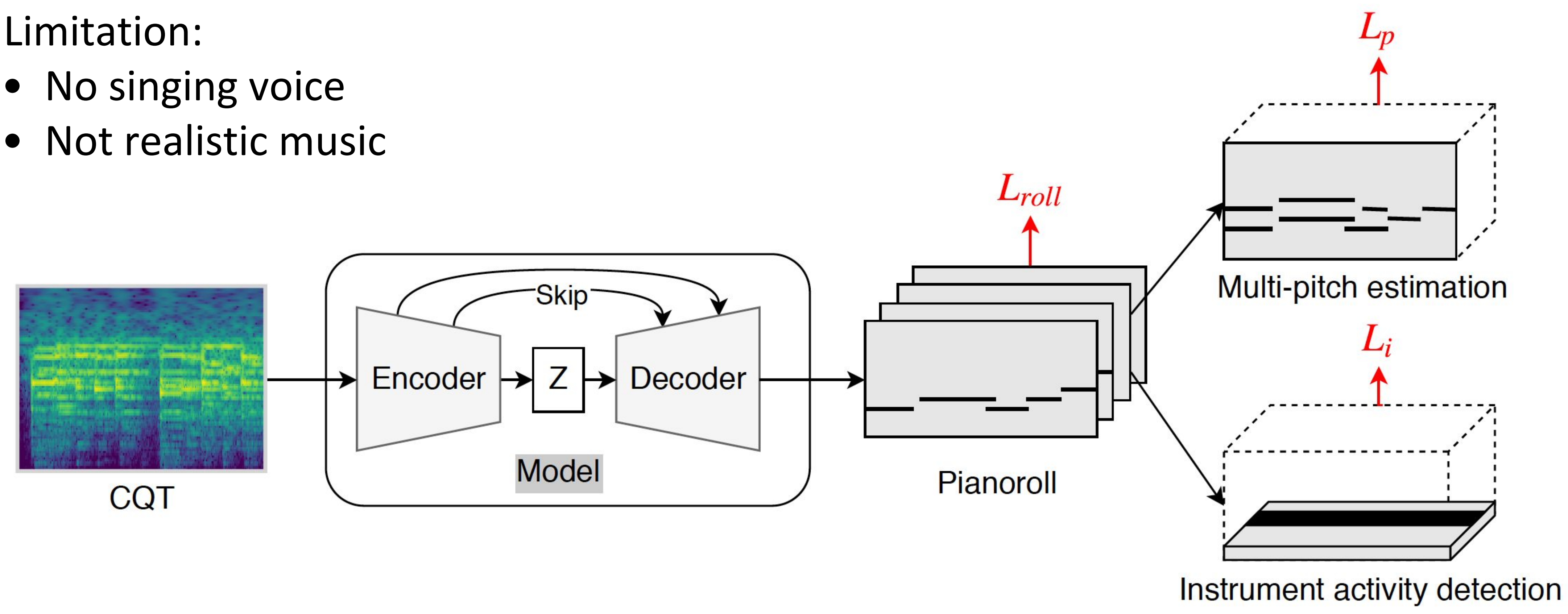
- Collect more than 344,166 pieces of song from Musescore forum
- Paired mp3 and MIDI files
- Include variety of genre and 128 instruments
- Synthesized music (from variety of synthesizers)
- We process the MIDi files to pianoroll, multi-pitch labels and instrument frame labels

Limitation:

- No singing voice
- Not realistic music

Dataset	Pitch labels	Instrument Labels	Real or Synth	Genre	Numbers of songs
MedleyDB	△ (partially)	✓	Real	Variety	122
MusicNet	✓	✓	Real	Classic	330
Bach10	✓	✓	Real	Classic	10
Mixing Secret		✓	Real	Variety	258
Musescore (in this paper)	✓	✓	Synth	Variety	344,166

System



- Unet as the main model structure
- The encoder and decoder are composed of four residual blocks. Each residual block has three convolution/up-convolution, two batchNorm and two leakyReLU layers.
- Binary Cross Entropy between ground truth and predicted value
- Doing three tasks at the same time:
  - o Piano roll prediction
  - o Multi-pitch estimation
  - o Instrument activity detection

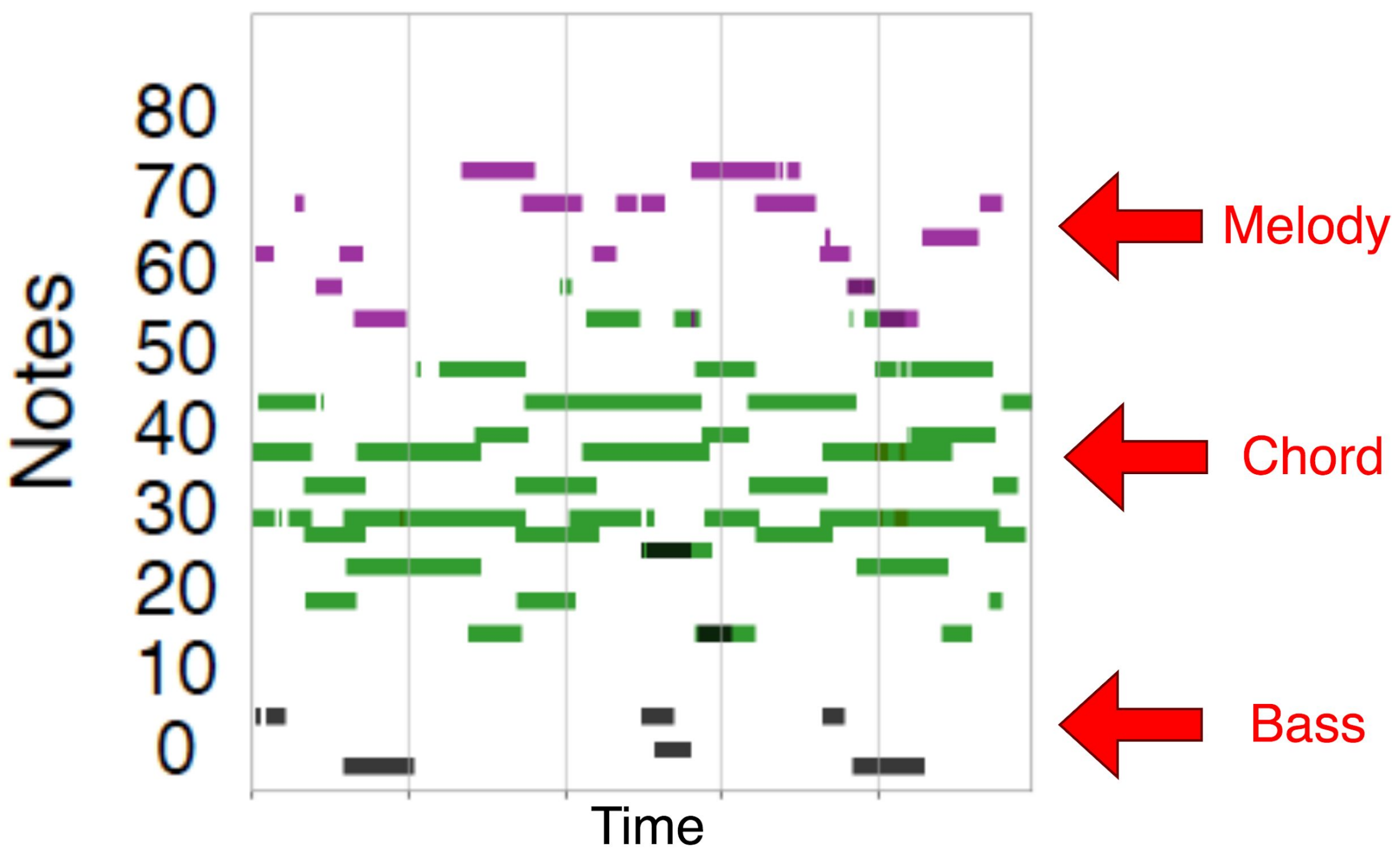
Result

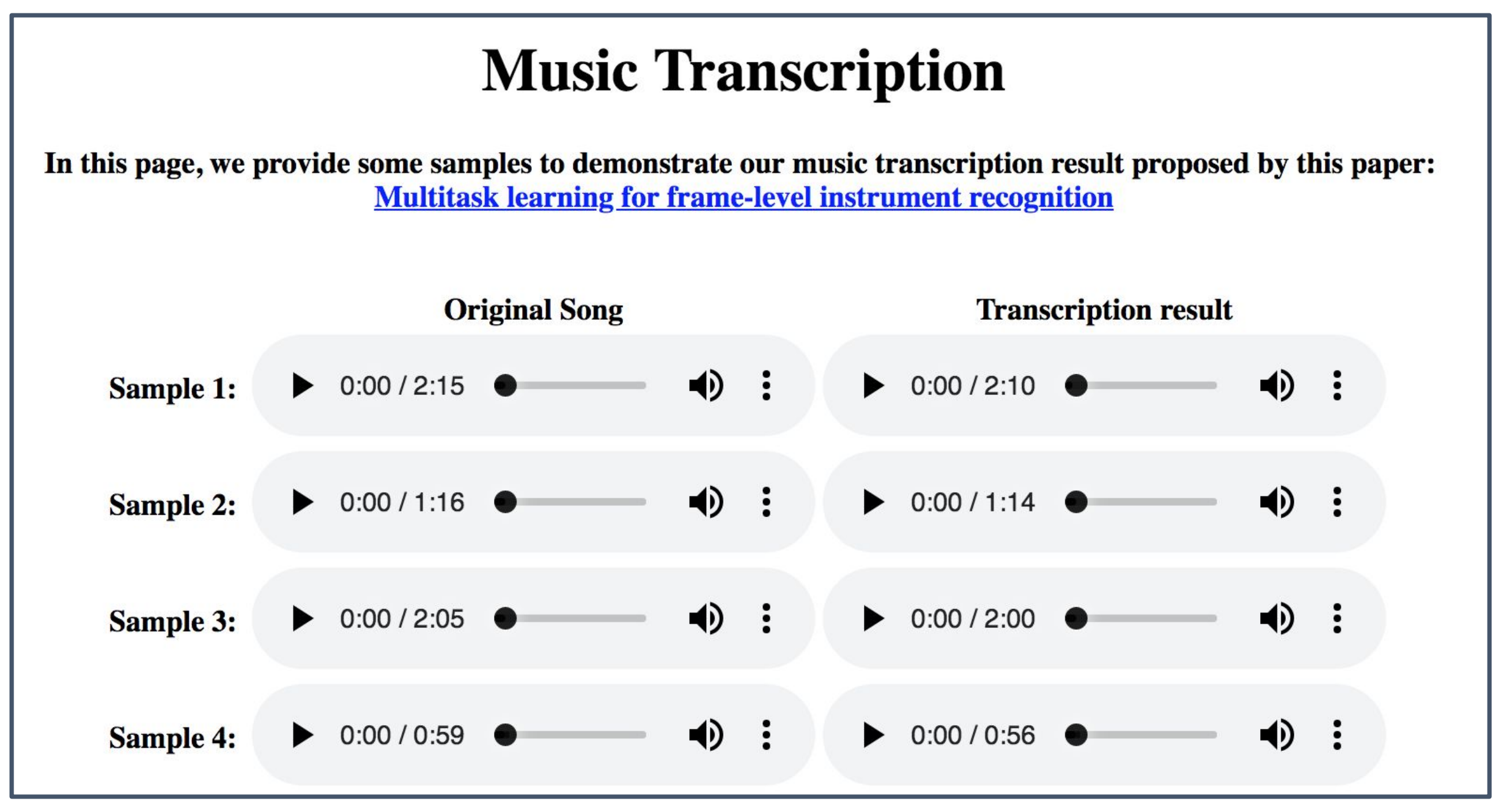
Method	Instrument	Pitch	Pianoroll
L_{roll} only (ablated)	—	—	0.623
L_i only (ablated)	0.896	—	—
L_p only (ablated)	—	0.799	—
all (proposed)	0.947	0.803	0.647

Method	Training Set	Piano	Guitar	Violin	Cello	Flute	Avg
[1]	YouTube-8M	0.766	0.780	0.787	0.755	0.708	0.759
[2]	Training split of 'MedleyDB+Mixing Secrets'	0.733	0.783	0.857	0.860	0.851	**0.817**
[3]	MuseScore training subset	0.690	0.660	0.697	0.774	0.860	0.736
Ours	MuseScore training subset	0.718	0.819	0.682	0.812	0.961	**0.798**

- Multitask learning is better than single task learning method

- Different methods but same testing set in [2]
- Testing set includes multi-instrument and singing voice
- F1-score of each instrument
- Compares favorably with [2]





Music Transcription

In this page, we provide some samples to demonstrate our music transcription result proposed by this paper:
Multitask learning for frame-level instrument recognition

	Original Song	Transcription result
Sample 1:	▶ 0:00 / 2:15	▶ 0:00 / 2:10
Sample 2:	▶ 0:00 / 1:16	▶ 0:00 / 1:14
Sample 3:	▶ 0:00 / 2:05	▶ 0:00 / 2:00
Sample 4:	▶ 0:00 / 0:59	▶ 0:00 / 0:56

Multi-pitch streaming overview!!

Future Work

- Using different synthesizers to augment our data
- Include singing voice into our model
- Increase instrument categories
- Music style transfer: change the latent vector Z in a meaningful way so that the output score can be modified too

Reference

[1] Jen-Yu Liu, Yi-Hsuan Yang, and Shyh-Kang Jeng,"Weakly-supervised visual instrument-playing action detection in videos," IEEE Trans. Multimedia , in press.
[2] Siddharth Gururani, Cameron Summers, and Alexander Lerch, "Instrument activity detection in polyphonic music using deep neural networks," in Proc. ISMIR, 2018.
[3] Yun-Ning Hung and Yi-Hsuan Yang, "Frame-level instrument recognition by timbre and pitch," in Proc. ISMIR ,2018, pp. 135–142.