Frame-level Instrument Recognition by Timbre and Pitch

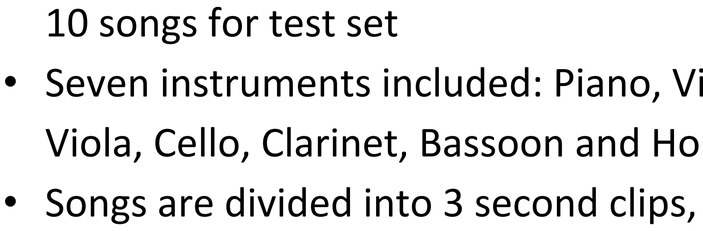
Yun-Ning Hung and Yi-Hsuan Yang (biboamy@citi.sinica.edu.tw)

Research Center for IT innovation, Academia Sinica, Taipei, Taiwan

[Demo Website] https://biboamy.github.io/instrument-recognition/demo.html

Introduction

- What is frame-level recognition?



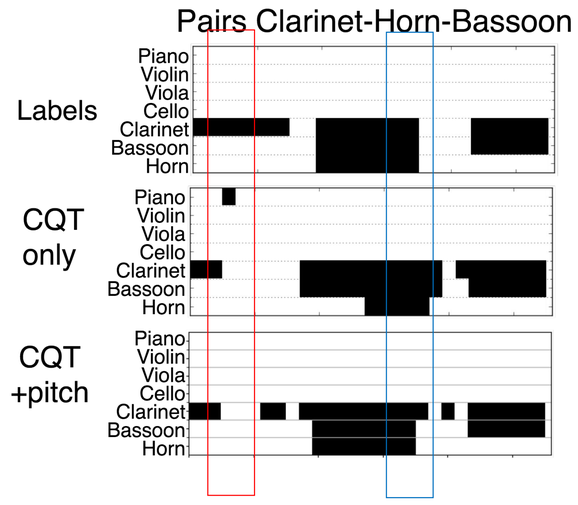
- Frame-level instrument prediction is important for music transcription, music structure analysis and other retrieval problems
- Very few datasets contain instrument frame labels

Dataset	Multi instrument	Instrument Frame Labels	Pitch Frame Labels
ParisTech, UIOWA, RWC			
IRMAS, AudioSet, MagnaTagATune	✓		
MedleyDB	✓	✓	Part of it
Mixing Secret	✓	✓	
MusicNet	✓	✓	✓

- Why MusicNet
  - Larger size
  - Pitch annotations

- Why we need pitch annotations?
  - Help catch onset/offset
  - Help catch harmonic distribution
  - Instrument chromatic range

- Network structure
  - Convolutional Neural Network
  - Preserve temporal dimension
  - Frame-label output

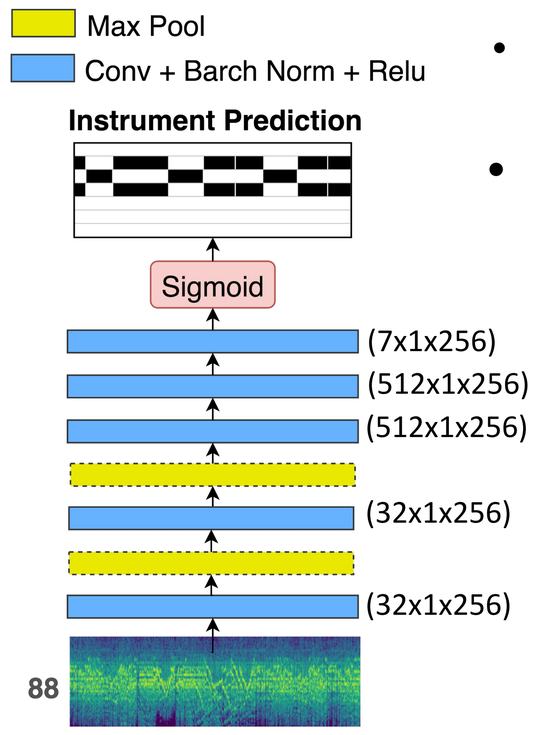


Data

MusicNet dataset [3]:

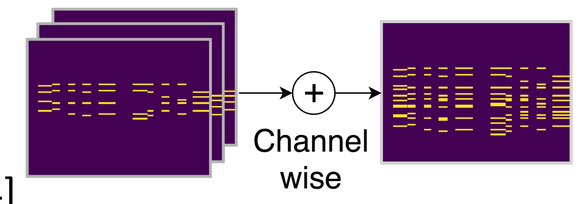
- 320 songs for training set and 10 songs for test set
- Seven instruments included: Piano, Violin, Viola, Cello, Clarinet, Bassoon and Horn
- Songs are divided into 3 second clips, with 88 (Piano notes) frequency bins and 258 time steps

System

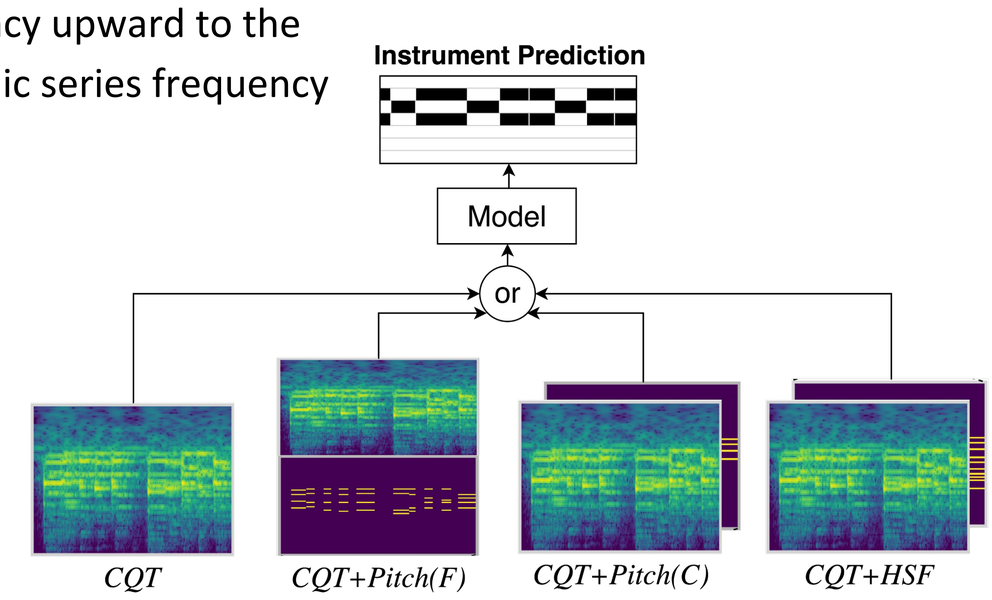


Input feature:

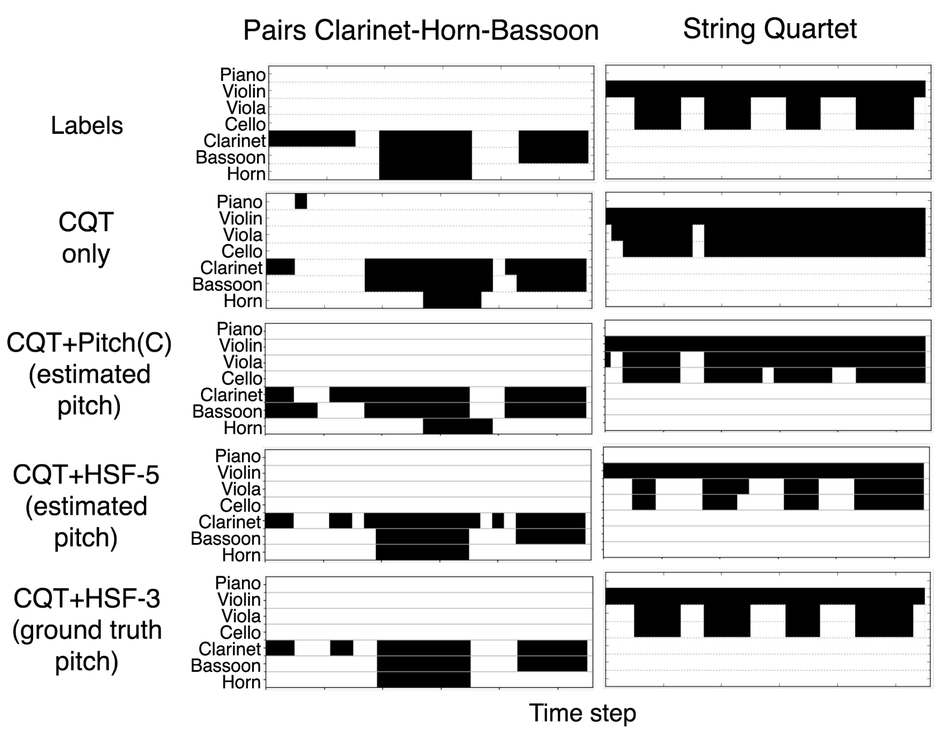
- Constant Q transform (CQT)
- Pitch: Ground truth pitch or pitch estimated by estimator [4]
- HSF: Shifting fundamental frequency upward to the harmonic series frequency



Harmonic Series Feature(HSF)



Result



Pitch Source	Method	Piano	Violin	Viola	Cello	Clarinet	Bassoon	Horn	Avg
none	CQT only	0.982	0.956	0.830	0.933	0.894	0.822	0.789	0.887
Estimated Pitch	CQT+Pitch(C)	0.982	0.958	0.819	0.921	0.898	0.827	0.794	0.886
	CQT+HSF	0.984	0.956	0.835	0.935	0.915	0.839	0.805	0.896
Ground Truth Pitch	CQT+HSF	0.997	0.985	0.914	0.971	0.944	0.907	0.810	0.933

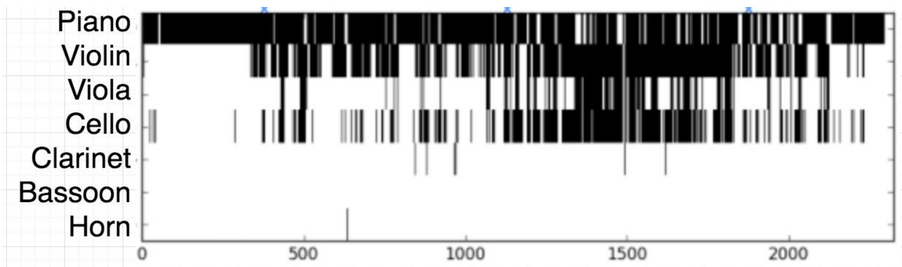
Adding pitch information can achieve higher F1-score than without pitch

Adding HSF can achieve higher F1-score than just adding pitch

We can reach 93% F1-score as the uper bound of adding pitch

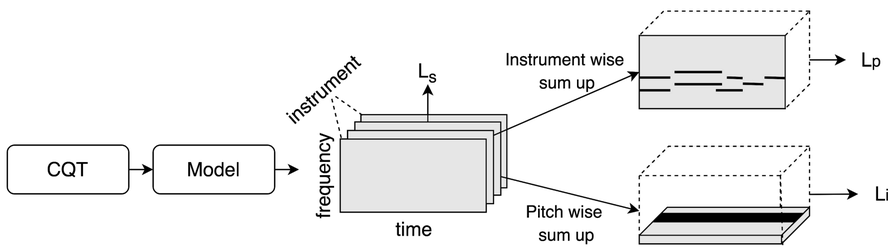
Improvement

- Instrument catagories: We will include MedleyDB & Mixing Secret in the future to cover more instrument and genre
- Temporal modeling: Include RNN to our network structure



Adele-Make you feel my love

Future Work



- Pitch and timbre joint learning
- Audio transcription
- Piano-rolls generation
- Pitch and timbre disentanglement
- Pitch and timbre representation learning
- Domain adaptation

Reference

[1] Liu et al. “Event localization in music auto-tagging.” *ACMMM 2016*.
[2] Chou et al. "Learning to Recognize Transient Sound Events using Attentional Supervision." *IJCAI*. 2018.
[3] https://homes.cs.washington.edu/~thickstn/musicnet.html
[4] Thickstun et al. Invariances and Data Augmentation for Supervised Music Transcription. ICASSP, 2018.