PiCoGen: Generate Piano Covers with a Two-stage Approach

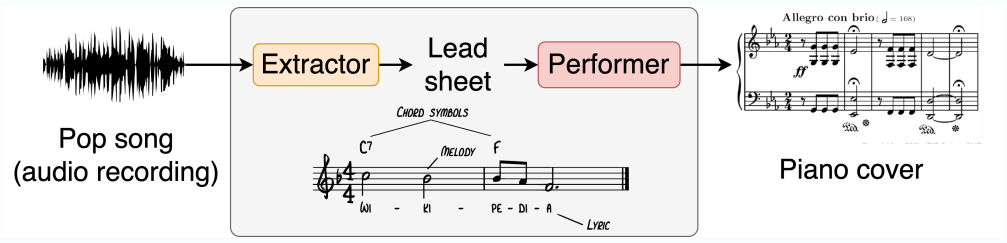
Chih-Pin Tan[1,2] Shuen-Huei Guan[2] Yi-Hsuan Yang[1]

[1]National Taiwan University [2]KKCompany Technologies

Abstract

We propose PiCoGen, a two-stage piano cover generation system

- ➡ *Stage1*: Transcribes the melody line and chord progression from audio recording
- ➡ *Stage2*: Generate a piano cover with the resulting lead sheet
- ➡ *No requirement of paired data* of covers and their original songs for training
- ➡ Generalizability across different musical genres



Dataset

Hook Theory:

- ~40k audio clips with labels of melodies and chords

Pop1k7:

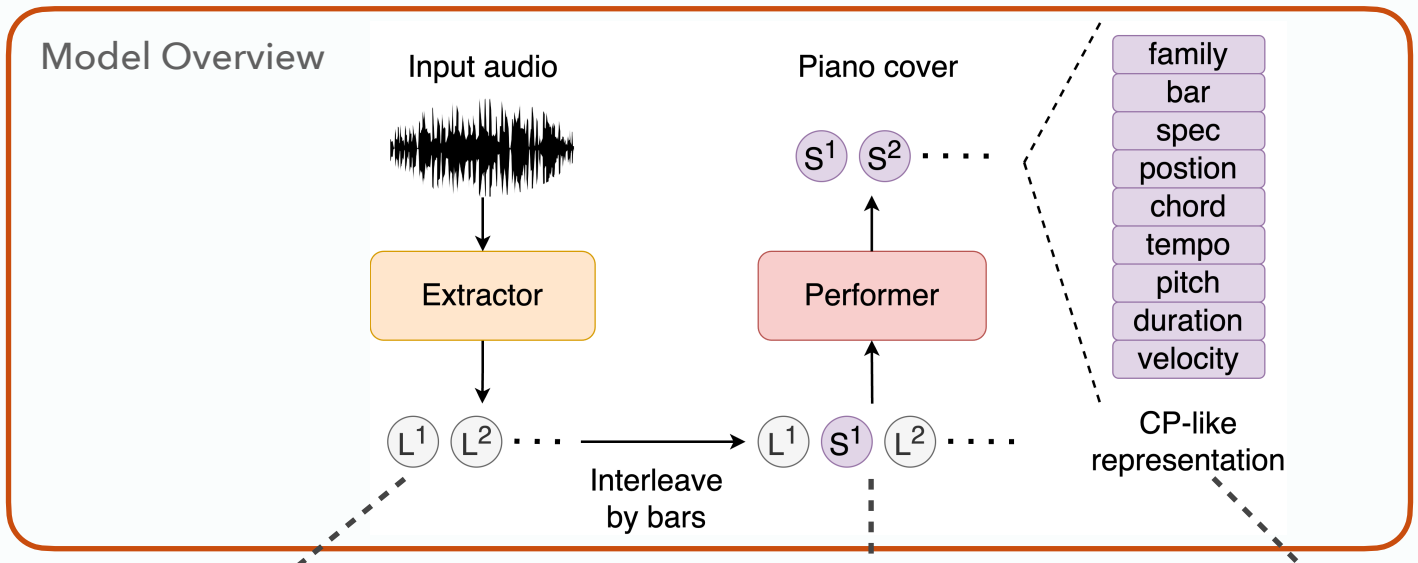
- ~1k7 piano performance of pop and anime songs

GTZAN:

- A collection of 10 genres with 100 audio files each, all having a length of 30 seconds.

Hook Theory dataset is not used in this paper since we build Extractor with the pre-trained model of SheetSage.

Architecture



Extractor

The stage 1 of PiCoGen is to convert the audio input into an intermediate representation.

We employ **SheetSage**, the stage-of-the-art lead sheet transcription model as the extractor to transcribe input audio to **human-readable lead sheet**

Performer

The stage 2 of PiCoGen is to generate the piano cover given on the extracted lead sheet which is conditional **independent** to the stage 1.

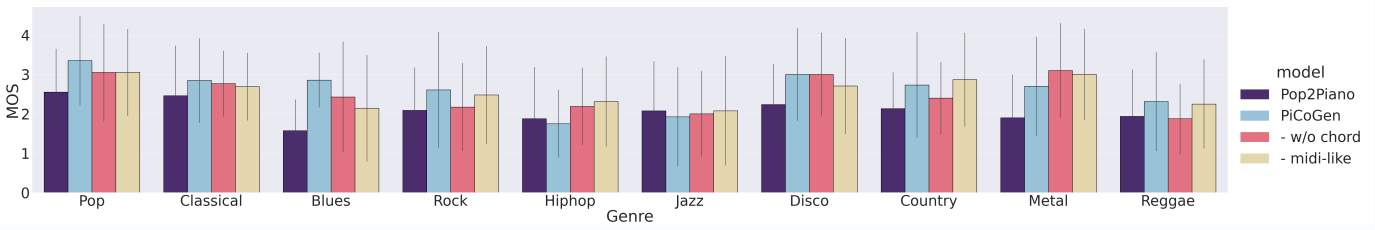
We implement the Performer with a **decoder-only Transformer** to deal with an interleaved sequence composed of a "bar-wise mix" of the condition, the lead sheet, and the output, the target piano cover.

Compound-Word

Among MIDI-compatible symbolic representations, compound-word (CP) token representation has two advantages: (i) lower complexity than **REMI** tokens, and (ii) easier to learn long-term dependency than **MIDI-like** tokens.

Experiment

User Study (OVL) in 10 Genres



Evaluation Results

Model	*objective* MCA ↑	*subjective evaluation* ↑ SI	SI_m	SI_c	FL	OVL
Pop2Piano[3]	**0.25**	2.65	2.85	2.65	2.60	2.55
PiCoGen	0.17	**3.20**	**3.35**	**3.05**	**3.20**	**3.35**
−*w/o chord*	0.14	3.10	3.25	2.95	3.10	3.05
−*midi-like*	0.12	2.95	3.10	2.95	3.10	3.05

- There are 2 ablations:
  1. Removing chord information from extracted lead sheet
  2. Using MIDI-like tokens instead of CP words

The experiment results show that PiCoGen outperforms on all aspects of the user study, including how the piano cover sounds similar to its original song (SI, SI_m, SI_c), how fluent the piano cover sounds (FL), and the overall assessment (OVL).

Conclusion

PiCoGen treats the lead sheet as the common ground between the input music and the target piano

- ➡ Bypass the need of curating paired data of covers and their original songs

The system performance is limited by accuracy of lead sheet extraction, which is one of the directions for future improvement.

https://tanchihpin0517.github.io/PiCoGen/

Check the demo website!