LEARNING TO GENERATE PIANO MUSIC WITH SUSTAIN PEDALS

Joann Ching[1,2]   **Yi-Hsuan Yang**[1,3]

[1] Research Center for IT Innovation, Academia Sinica
[2] Center for Music Technology, Georgia Institute of Technology, USA
[3] Yating Music Team, Taiwan AI Labs
`jching9@gatech.edu, yang@citi.sinica.edu.tw`

ABSTRACT

Recent years have witnessed a growing interest in research related to the detection of piano pedals from audio signals in the music information retrieval community [1–4]. However, to our best knowledge, recent generative models for symbolic music have rarely taken piano pedals into account. In this work, we employ the transcription model proposed by Kong *et al.* [4] to get pedal information from the audio recordings of piano performance in the AILabs1k7 dataset [5], and then modify the Compound Word Transformer proposed by Hsiao *et al.* [5] to build a Transformer decoder that generates pedal-related tokens along with other musical tokens. While the work is done by using inferred sustain pedal information as training data, the result shows hope for further improvement and the importance of the involvement of sustain pedal in tasks of piano performance generations.

1. INTRODUCTION

Research on automatic music generation dates back to the 80s. With the trend in recent days, results of such models have improved vastly [5–7]. However, while the usage of sustain pedal is a common practice in piano performances nowadays, there has not been much work that includes such elements to recent deep learning-based generation models, possibly due to the scarcity of data. Piano performances are comprised of not only compositional features such as pitch and duration but also performance features. Given the relatively mature work of automatic music generation with compositional features, we want to investigate adding piano pedals to generative models of music.

Such features are difficult for a model to learn because of its subjective nature being based on a performer's interpretation, especially in classical piano music. Within the broad idea of classical music, there are sub-genres such as Baroque, Classical, and Romantic, all possessing their style of playing, not to mention the usage of pedaling.
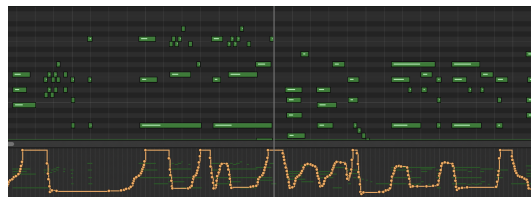
(a) Visualization of a midi file from the MAESTRO dataset
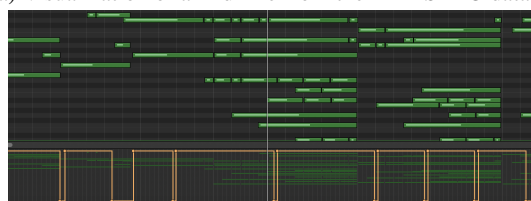
(b) Visualization of a midi file from the AILabs1k7 dataset

Figure 1: Illustration of the usage of the sustain pedal (orange lines) in MAESTRO [8] (classical; ground truth pedal labels) and AILabs1k7 [5] (pop, pseudo labels).

For example, techniques of using the sustain pedal, especially in classical music, can be categorized into three types: *Anticipatory*, *Rhythmic*, and *Legato*. In current existing symbolic datasets, MAESTRO [8] is the only one that contains piano pedal information collected from real performances. However, the usage difference between the three types of techniques might be too vague for a machine learning model to distinguish.

Although pedaling is arguably more important in classical music, we choose to start with using pop music as training data, given the simplicity of pedal techniques in pop music (e.g., pedal changes have a higher relationship with both beat and chord; see Fig. 1). There are three types of pedals on a piano, the Soft Pedal, the Sostenuto Pedal, and the Sustain Pedal, but we only focus on the sustain pedal in this work. The source code for implementing the generation model can be found at a GitHub repo. [1] Examples of generated pieces can be found at a demo webpage. [2]

2. DATASET

We use the AILabs1k7 dataset compiled by Hsiao *et al.* [5] for training. The dataset contains 1,748 audio sam-
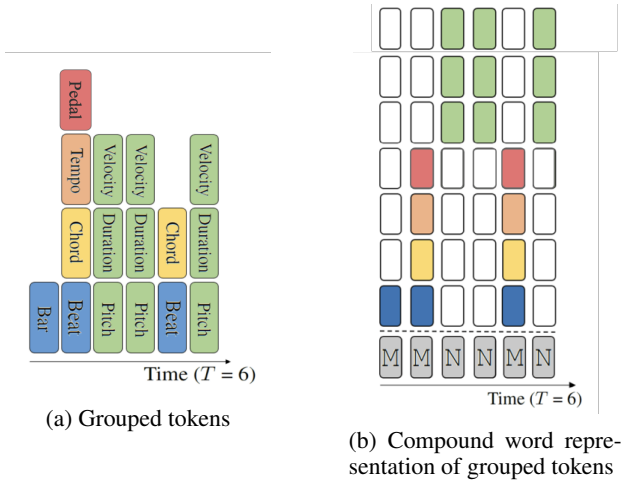
---

[1] `https://github.com/joann8512/SusPedal-Gen`
[2] `https://joann8512.github.io/SusPedal-Gen/`

(a) Grouped tokens

(b) Compound word representation of grouped tokens

Figure 2: Example visualization of tokens

	Midi-like [6]	Ours
Note onset	Note-On (0-127)	Note-On (0-127)
Note offset	Note-Off (0-127)	Note-Duration (1–64 32nd notes)
Time grid	Time-shift (10–1000ms)	Subbeat (16 bins) & Bar (1 kind)
Tempo changes	—	Tempo (56 kinds)
Chord	—	Chord (60 types)
Pedal	Control Change (2 kinds)	Pedal (10 types)

Table 1: Representation comparison on the "MIDI-like" event representation used by [6] and our representation. Sizes of the token types are shown in brackets.

ples of pop piano cover music collected from the Internet, all of which are in 4/4 time signature (i.e., 4 quarter notes in each bar). We then transcribed the audio files into symbolic sequences using the open-source state-of-the-art high-resolution piano transcription model proposed by Kong *et al.* [4] for its ability to transcribe sustain pedal information. It is possible to have undesirable cases from the transcription results for some of the songs, given that Kong *et al.*'s model was trained for transcribing classical piano music. For example, notes are fragmented when the recording is engineered to have too many ambient effects (e.g., some songs in the EMOPIA dataset [7], according to our preliminary experiments). For AILabs1k7, the transcription result for pedals is acceptable.

3. DATA REPRESENTATION

To train a Transformer model for generation, we need a symbolic data representation that can be used as input to the model. The encoded symbolic data from the dataset are represented as "event tokens" to the model. Under the Compound Word Transformer, tokens with similar musical meaning are grouped to the same *family*, which are then grouped into a *super token* and placed on the same timestamp. The family groups contains *metrical*, *note*, and *end of sequence*. In pop music, the most common and easiest way of pedaling is to change together with chord changes. Therefore, we place the pedal tokens in the metrical family group, along with the chord and tempo tokens, as depicted in Figure 2b. The presence of a pedal token signifies the onset time of using the sustain pedal.

The value taken by a pedal token indicates the pedal duration. We quantize the pedal duration and consider only the following 10 types of duration: 240, 480, 720, 960, 1440, 1920, 2400, 2880, 3360, 3840 ticks, where one beat corresponds to 480 ticks. The shortest is then an eighth note and the longest is 8 beats. We apply quantization given that the transcribed pedal duration is not 100% accurate anyway, and that reducing the vocabulary size of the pedal tokens (i.e., to 10) makes it easier for Transformers to learn the relations between notes and pedals.

4. FUTURE WORKS

In the current stage of the work, it is only proven that the results are acceptable. If the generation quality can be improved, with the sustain pedal information included in the generation process, the musical effects can be greatly improved in automatically generated music. For example, different emotions in music are often expressed through pedaling. In other words, pedaling in music generation could help to show the difference of emotions, and thereby creating more diversity in generated music.

5. REFERENCES

[1] B. Liang, G. Fazekas, and M. Sandler, "Detection of piano pedaling techniques on the sustain pedal," *Journal of the Audio Engineering Society*, 2017.

[2] ——, "Piano legato-pedal onset detection based on a sympathetic resonance measure," in *Proc. EUSIPCO*, 2018.

[3] ——, "Transfer learning for piano sustain-pedal detection," in *Proc. IJCNN*, 2019.

[4] Q. Kong *et al.*, "High-resolution piano transcription with pedals by regressing onsets and offsets times," *arXiv preprint arXiv:2010.01815*, 2020.

[5] W.-Y. Hsiao *et al.*, "Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proc. AAAI*, 2021.

[6] C.-Z. A. Huang *et al.*, "Music Transformer," in *Proc. ICLR*, 2019.

[7] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *Proc. ISMIR*, 2021.

[8] C. Hawthorne *et al.*, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. ICLR*, 2019.