SCORE TRANSFORMER: TRANSCRIBING QUANTIZED MIDI INTO COMPREHENSIVE MUSICAL SCORE

Masahiro Suzuki
Yamaha Corp.
masahiro1.suzuki@music.yamaha.com

ABSTRACT

We explore the tokenized representation of musical scores to generate musical scores with transformers. We design score token representation corresponding to the musical symbols and attributes used in musical scores and train the Transformer model to transcribe note-level representation into musical scores. Evaluations of popular piano scores show that our model significantly outperforms existing methods on all 4 investigated categories. We also explore an effective token representation, including those based on existing text-like score formats, and show that our proposed representation produces the steadiest results.

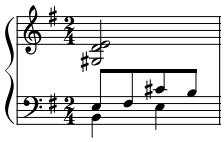
1. INTRODUCTION

Deep neural networks (DNNs) have yielded impressive results in music generation and music transcription. However, their application to the generation of a comprehensive musical score or even its effective representations remains unexplored. In music transcription, for example, DNNs have achieved remarkable success in the audio-to-MIDI process (e.g., multi-pitch estimation and onset/offset detection) [1]. In contrast, the subsequent MIDI-to-score process, in which note-level representation is transcribed into music notation [2], has not been comprehensively addressed in prior studies [3]. Specifically, there has been a lack of focus on the musical score generation (or score typesetting) subtask, in which quantized MIDI is transcribed into musical scores. In this work, we address the generation of comprehensive musical scores using the Transformer model [4], focusing on piano scores.

2. SCORE TOKENIZATION

We design a token representation that symbolizes score elements. Our design principles are as follows: 1) one token per score symbol or note attribute, 2) compatible with `music21` [5] attributes to build scores easily, 3) concatenated sequence of staves to generate multi-stave scores consistently, and 4) tokenize essential symbols excluding expression and repeat symbols.

Table 1 lists the symbols used in this work. Figure 1(c) shows an example score token sequence. We tokenize score symbols and attributes in each staff from left to right



(a) Score

```
R bar clef_treble key_sharp_1 time_2/4
note_E4 note_D4 note_G#3 len_2 stem_up bar
… L bar clef_bass key_sharp_1 time 2/4
<voice> note_E3 len_1/2 stem_up beam_start
note_F#3 len_1/2 stem_up beam_continue
note_C#4 stem_up beam_continue note_B3
len_1/2 stem_up beam_stop </voice> <voice>
note_B2 len_1 stem_down note_E3 len_1
stem_down </voice> bar …
```

(b) Notation-level Representation

```
bar note_64 len_48 note_62 len_48 note_56
len_48 note_52 len_12 note_47 len_24
pos_12 note_54 len_12 beat note_61 len_12
note_52 len_24 pos_12 note_59 len_24 bar
…
```

(c) Note-level Representation

Figure 1. Example of our token representations corresponding to (a) score excerpt.

Symbol	Example	Variations
Staff	`R`	`R/L`
Barline	`bar`	`bar`
Clef	`clef_treble`	`clef_{bass/treble}`
Key Signature	`key_flat_2`	`key_{sharp/flat/natural}_{1, 2, …, 6}`
Time Signature	`time_4/4`	`time_{2/4, 3/4, 4/4, etc.}`
Voice	`<voice>`	`<voice>, </voice>`
Rest	`rest`	`rest`
Pitch	`note_C4`	`note_{A, B, ..., G}{##/#/b/bb/(none)}{0, 1, …, 8}`
Duration	`len_1/2`	`len_{1/24, 1/16, …, 4}`
Stem Direction	`stem_up`	`stem_{up/down}`
Beams	`beam_stop`	`beam_{start/stop/continue/partial-left/partial-right}_...`
Tie	`tie_start`	`tie_{start/continue/stop}`

Table 1. Symbols and their variations in the proposed score token representation.

with an exception for voices, where pairs of tag-like tokens indicate concurrent voices. We concatenate the token sequences of the *right* and *left* hands with the marking tokens `R` and `L` (Fig.1(b)). Whereas rest representation always consists of *Rest* and *Duration* tokens, the note representation consists of up to 5 types of tokens (from *Pitch* through *Tie* in Table 1). Chords are expressed as consecutive *Pitch* tokens (e.g., the second line in Fig.1(b)).

3. FROM MIDI TO SCORE

We train the model using musical score data to restore the original score from down-converted single-track MIDI

data. For note-level (e.g., MIDI) tokenization, we adopt REMI [6], expanding it with the `beat` token to deal with various meters. An example of note-level tokenization is shown in Fig. 1(c). The model is trained to convert the *note*-level sequences into *notation*-level sequences (e.g., Fig. 1(b)).

4. EXPERIMENTAL SETTINGS

Dataset. We employed 2,863 popular piano scores for the experiments. The scores were split system by system and tokenized into token sequences. We split the dataset 8:1:1 song-wise for training, validation, and test, respectively.

Model. We used small setting for the Transformer model with approximately 5M parameters ($d_{\text{model}} = 256$, $d_{\text{ff}} = 512$, $h = 4$, and $N = 3$ with the same definition as in [4]).

Baselines. We adopted the music transcription framework proposed in [3] as a baseline, denoting it as "CTD." We also employed Finale 26 and MuseScore 3 as baselines.

Alternative tokenization methods. We also adopted existing text-like score formats (e.g., ABC, Humdrum [7], and LilyPond [8]) and created token sequences by segmenting the score-formatted strings. For Humdrum, we tokenize its 2D representation on a row-major or a spine-major basis, denoting them as *row* and *spine*, respectively.

Metric. We used the metric proposed in [9] that measures music notation quality based on the number of errors (vs. ground truth score) on 12 musical aspects of a score. We included 3 musical aspects (*voice*, *beam*, and *tie*) while excluding two redundant aspects (*barline* and *note grouping*). We aggregate the aspects into 4 categories.[1]

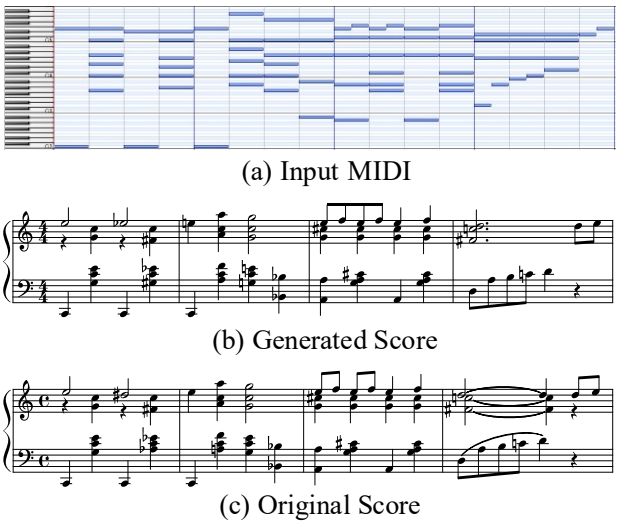
5. RESULTS AND ANALYSES

	Note Preservation	Note Segregation	Score Attributes	Note Attributes	Average
ST	**2.96**	**1.88**	4.59	**2.06**	**2.81**
w/ABC	15.41	5.00	4.53	9.43	8.38
w/Humdrum(row)	5.21	5.21	6.26	4.17	5.04
w/Humdrum(spine)	5.95	3.85	6.88	4.47	5.22
w/LilyPond	3.48	6.00	**4.17**	2.38	3.61
CTD [3]	97.29	42.18	17.97	32.11	41.93
Finale 26	28.79	18.92	13.95	15.60	17.94
MuseScore 3	13.43	31.72	16.96	11.67	16.63

Table 2. Overall error rates in % (measured based on the difference between the original and generated scores).

5.1 Comparison with Baselines

Table 2 shows the overall results measured in error rates for 4 categories. According to these results, the Score Transformer (ST) performed significantly better than the baseline methods (lower half) on all 4 categories. The results suggest that ST not only successfully learned how to transcribe note-level representation into a musical notation, but also has a much higher capability to notate music than those of prior arts. The results also demonstrate that ST can jointly estimate various symbols and attributes in musical scores. Figure 2 shows an example of the generated musical scores. Although minor differences can be observed when compared to the original score, ST succeeded in generating an appropriate score out of input MIDI.



(a) Input MIDI

(b) Generated Score

(c) Original Score

Figure 2. Example of generated score.

5.2 Evaluation of Alternative Tokenization

By comparing the error rates of Score Transformer (ST) and its variants with those of existing formats in Table 2 (see upper half), we can observe that the methods that use existing formats exhibit unstable (variably low to rather high) error rates among the investigated categories. By contrast, the proposed tokenization method (ST) demonstrated a stable performance over these categories. The result suggests that the proposed tokenization method produced the steadiest results with the Transformer model.

6. CONCLUSION

We designed score tokens to represent musical scores and trained the Transformer model using paired sequences of *note*-level and *notation*-level data. The evaluation results show that our model significantly outperformed the baseline methods. Additionally, we demonstrated that our score token representation is among the most effective via a comparison with various tokenization methods. The tools to convert between the representation and MusicXML are provided.[2] We believe that our method opens new possibilities for research into a variety of tasks that involve musical scores.

[1] The mapping between the 4 categories and aspects are as follows:
[Note Preservation] *note* and *rest*.
[Note Segregation] *staff assignment* and *voice separation*.
[Score Attributes] *clef*, *key signature* and *time signature*.
[Note Attributes] *duration*, *spelling*, *stem direction*, *beam* and *tie*.

[2] https://github.com/suzuqn/ScoreTransformer

7. REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert. Automatic music transcription: an overview, *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[2] F. Foscarin, F. Jacquemard, P. Rigaux, and M. Sakai. A parse-based framework for coupled rhythm quantization and score structuring. In *Proceedings of the International Conference on Mathematics and Computation in Music*, 2019, pp. 248–260.

[3] A. Cogliati, D. Temperley, and Z. Duan. Transcribing human piano performances into music notation. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, New York, USA, 2016, pp. 758–764.

[4] A. Vaswani *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[5] M. S. Cuthbert and C. Ariza. music21: a toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, 2010, pp. 637–642.

[6] Y.-S. Huang and Y.-H. Yang. Pop music transformer: generating music with rhythm and harmony. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[7] D. Huron. Music information processing using the humdrum toolkit: concepts, examples, and lessons. *Computer Music Journal*, vol. 26, no. 2, p. 11, 2002.

[8] H.-W. Nienhuys and J. Nieuwenhuizen. Lilypond, a system for automated music engraving. In *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003)*, Firenze, Italy, 2003, pp. 1–6.

[9] A. Cogliati and Z. Duan. A metric for music notation transcription accuracy. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017, pp. 407–413.