PIANO FINGERING ESTIMATION AND COMPLETION WITH TRANSFORMERS

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

ABSTRACT

We formulate fingering-related tasks as sequence-to-sequence problems and solve them with the Transformer model. By integrating finger information into REMI-based tokens, we show our trained model can 1) estimate partial fingering (fingering is estimated for only partial notes), and 2) complete fingering from partially specified fingering.

1. INTRODUCTION

Fingering information is useful for playing and practicing keyboard instruments. For beginner players, fingering is a necessity and should be indicated on all the notes on scores. In contrast, for intermediate players, fingering is usually sufficient only for partial notes. Figure 1 illustrates three possible fingering states of scores and fingering-related tasks between them. Although automatic fingering estimation has long been studied [1, 2], prior works have addressed only the complete fingering estimation task, which estimates fingering for all the notes in scores (Task 2 in Fig. 1). In this work, we extend the task to two tasks: the partial fingering estimation task (Task 1) and the fingering completion task (Task 3). The former task estimates and display fingering on partial notes, like intermediate scores. The latter task complete fingering from partially designated fingering, which may help beginners to tackle scores with higher level. We formulate these tasks as sequence-to-sequence problems and address them with the vanilla Transformer [3], a typical sequence model.

2. TOKEN REPRESENTATION

We adopt REMI-based tokens to efficiently represent notes, expanding REMI [4] to represent both note pitch and its fingering in a single token. We use simple token concatenation to express multiple properties allocated to a token, unlike advanced tokenization methods such as CP [5] and OctupleMIDI [6] incorporating embedding concatenation. Figure 2 shows the example of our token representation. Our representation consists of following 6 types of token: hand (`R/L`), barline (`bar`), note timings within bar (`pos`), note pitch (`note`), note duration (`len`) and fingering rate (`fingerrate`) tokens. For example, `note_60_1` indicates the note pitch in note number (i.e., 60 = C4) and its fingering (i.e., 1 = thumb); `note_62`, which has no fingering part in token, symbolizes the D4 note without fingering indication. We split each measure into 48 steps (regardless of beats per measure) and quantize note timings and durations with the steps. We condition the model with the target rate of notes with fingering. Following 5 ranges of rate are employed for conditioning: [0-12.5), [12.5-25), [25-50), [50-100), and 100%.

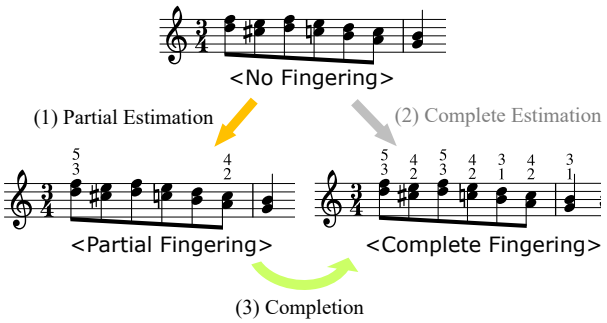
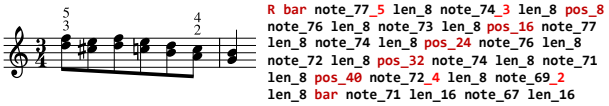


Figure 1. Overview of fingering tasks. Extending conventional (2) complete estimation task, we address (1) partial estimation and (3) completion tasks in this work.



(a) Score Excerpt (b) Token Representation

Figure 2. (b) Example of our token representation corresponding to (a) score excerpt. Fingering parts in `note` tokens are colored in red.

3. DATASET

We experimented using the scanned piano scores with fingering, consisting of pop and classical pieces (2,247 scores in total). They were split 8:2 score-wise for training and test, respectively. The scores were segmented 6-8 measures each (allowing overlap) and tokenized into sequences. We trained a separate model for each task. While all the resulting 182,060 segments were employed in the

fingering estimation task, the fingering completion task used only the segments that had 25% or more notes with fingering.

4. PARTIAL FINGERING ESTIMATION

4.1 Training

We trained the fingering estimation model with pairs of token sequences: sequences with fingering information removed (as *sources*) and the original sequences (as *targets*). We trained the model to restore the latter from the former, conditioning with the range of fingering rates of the latter.

4.2 Evaluation

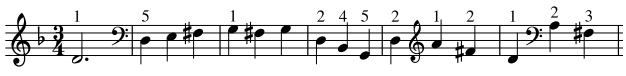
Fingering Position. Assuming that which note to have fingering in the original score is the only correct answer, Precision, Recall, and F1-score for model output were 0.71, 0.63, and 0.66 respectively.

Fingering Accuracy. Accuracy of the estimated finger for the fingering where estimated fingering positions were the same as original scores was 70.9%.

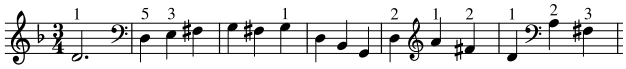
Input-Output Consistency. Sequence-to-sequence attribute estimation may result in inconsistent input and output. Comparing input and output notes, 99.82% of the inferred sequences had no inconsistency, suggesting disagreements of notes rarely occured with our token representation.

4.3 Example

Figure 3 demonstrates that the trained model successfully estimated playable fingering in a partial form. Comparing estimated and original fingering, although there are differences in the notes where fingering are displayed, the two fingerings can be considered almost practically the same. The result implies the correct answer of which notes to have fingering is ambiguous and it makes the training and evaluation of partial fingering estimation rather hard.



(a) Estimated Fingerings



(b) Original Fingerings

Figure 3. Example of partial fingering estimation (for left hand) for an excerpt of a classical piece.

5. FINGERING COMPLETION

5.1 Training

We trained the fingering completion model with the following pairs of token sequences: sequences with fingering information randomly dropped (as *sources*) and the original sequences (as *targets*). Fingering dropping rate was randomly picked from range 20-80% for each sequence.

5.2 Evaluation

Fingering Accuracy. Accuracy of the completed fingering (restoring the original fingering) from randomly dropped fingering were 83.6%.

Fingering Preservation. Specified fingering should remain the same before and after completion. Note-wise preservation rate of specified fingering was 99.98%, indicating the specified fingering were successfully preserved in almost all the inferred sequences.

5.3 Example

Figure 4 shows the results of fingering completion on different fingering specifications (only red-colored fingering were specified on the input sequence). The trained model seems to succeed in completing fingering naturally, taking into account the partial input fingering, even if the specified fingering were rather unusual (Fig. 4(d)). The result suggests the model may have the potential to complement the fingering in the score while respecting the designated fingering instruction. Moreover, fingering completion could be the aid to generate the personalized fingering with minimum effort, because fingering styles are known to be quite different between individuals.

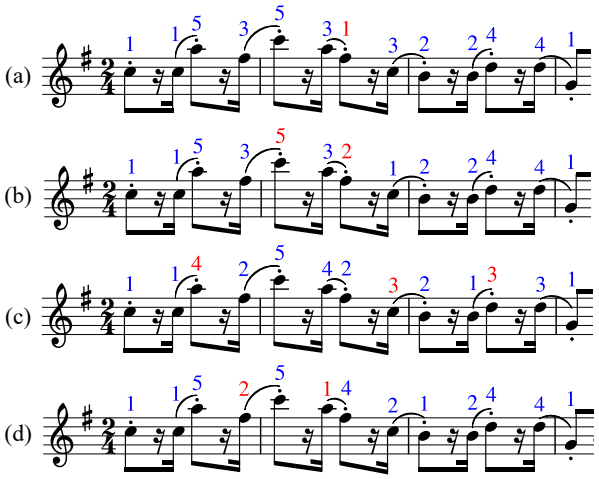


Figure 4. Example of fingering completion. We designated only the fingering in red. The model completed fingering by estimating the fingering for other notes.

6. CONCLUSION

We showed that the Transformer model trained with our fingering tokenization can estimate and complete fingering, which suggests sequence-to-sequence learning can be employed to solve fingering-related tasks. We also found that simple token concatenation without embedding operation can be used in such trainings. We plan to research these methods more extensively.

7. REFERENCES

[1] Y. Yonebayashi, H. Kameoka, and S. Sagayama, "Automatic decision of piano fingering based on hidden Markov models," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007, pp. 2915–2921.

[2] E. Nakamura, Y. Saito, and K. Yoshii, "Statistical learning and estimation of piano fingering," *Information Sciences*, vol. 517, pp. 68–85, May 2020.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, New York, USA, 2017, pp. 6000–6010.

[4] 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*, Feb 2020.

[5] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, "Compound word transformer: learning to compose full-Song music over dynamic directed hypergraphs," *arXiv preprint arXiv:2101.02402*, 2021.

[6] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, "MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training," *arXiv preprint arXiv:2106.05630*, Jun 2021.