MUSIC SCORE EXPANSION WITH VARIABLE-LENGTH INFILLING

Chih-Pin Tan[1] **Chin-Jui Chang**[2] **Alvin W.Y. Su**[1] **Yi-Hsuan Yang**[2,3]

[1] Department of Computer Science, National Cheng Kung University
[2] Research Center for IT Innovation, Academia Sinica
[3] Yating Music Team, Taiwan AI Labs

p76091551@gs.ncku.edu.tw, yhyang@ailabs.tw

ABSTRACT

In this paper, we investigate using the *variable-length infilling (VLI)* model [1], which is originally proposed to infill missing segments, to "prolong" existing musical segments at musical boundaries. Specifically, as a case study, we expand 20 musical segments from 12 bars to 16 bars, and examine the degree to which the VLI model preserves musical boundaries in the expanded results using a few objective metrics, including the *Register Histogram Similarity* we newly propose. The results show that the VLI model has the potential to address the expansion task.

1. INTRODUCTION

Music score infilling is an instance of automatic symbolic music generation tasks. Given two *disconnected* musical segments, an infilling model aims to "fill the gap" between them by generating novel content, as depicted in Figure 1. We refer to the two given segments as the *past context* C_{past} and the *future context* C_{future}, respectively, and the generated one as the *infilled segment* C_{new}. We assume that these segments are all represented by sequences of event tokens such as note-on and note-duration [2].

Very recently, we proposed a *variable-length infilling (VLI)* model [1] that accepts variable-length "context gaps" and generates variable-length musical contents. Specifically, at inference time, the input needed by the VLI model is composed of only the two token sequences C_{past}, C_{future}, and *the number of bars* between C_{past} and C_{future} (i.e., the context gap). VLI decides on its own the number of tokens to be filled to the designated gap between the given contexts. Moreover, a single VLI model can deal with different context gaps (instead of needing one model per context gap). In VLI, the context gap is expected to be a non-zero and positive integer.

Interestingly, we later realize that VLI also holds the potential to address a highly relevant, yet much less explored task, called **music score expansion**. Given a sequence of music notes $\{n_1, n_2, ..., n_L\}$, this task entails increasing the length of the sequence by *inserting* novel content at one or multiple locations of the sequence. Without loss of generalizability, we assume that we are about to insert content at only one location p, where $1 < p < L$. We can then use VLI to address this task by treating $\{n_1, n_2, ..., n_p\}$ as C_{past}, $\{n_{p+1}, n_{p+2}, ..., n_L\}$ as C_{future}, and setting the desired context gap to a non-zero integer. In other words, while the C_{past} and C_{future} are originally *connected* with zero gap in between, the idea here is to split them apart somewhere, create artificial gap between then, and use VLI to fill the gap, to prolong the sequence as a result. [1]

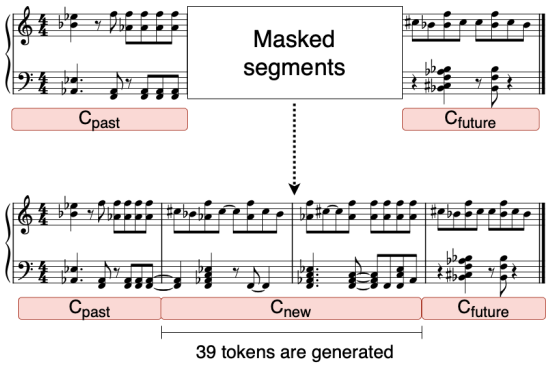

Figure 1. An illustration of how the variable-length infilling model [1] works, where the model is asked to generate 2-bar musical content (i.e., the "context gap" is two bars). We do not explicitly consider key in this work.

In this paper, we present preliminary experiments exploring such a novel use case of VLI, treating music score expansion as a sub-task of score infilling.

2. EXPERIMENT

While there are many ways to choose the location p for content insertion, in our work, we utilize the VLI model to expand short-length musical segments at *musical boundaries*. A musical boundary is viewed as the edge of two musical groups [3], which is often accompanied by changes in rhythm, metre and pitch patterns. It should be expected in general that there is still a musical boundary after expansion. This is the focus of our experiment.

[1] We can also "shorten" a sequence by removing a sub-sequence and use VLI to create a novel content that is shorter than the removed subsequence to connect the gap smoothly; we leave this as a future work.

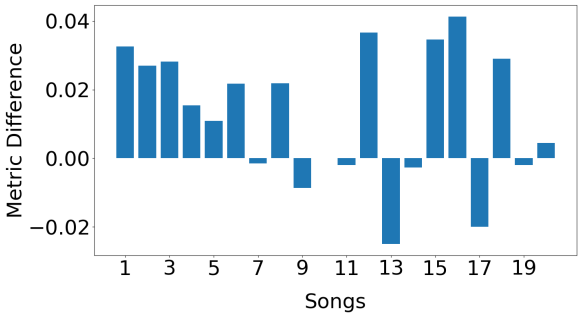

Figure 2. The result of subtracting the grooving pattern similarity of (C_{past}, C_{new}) from that of (C_{new}, C_{future}).

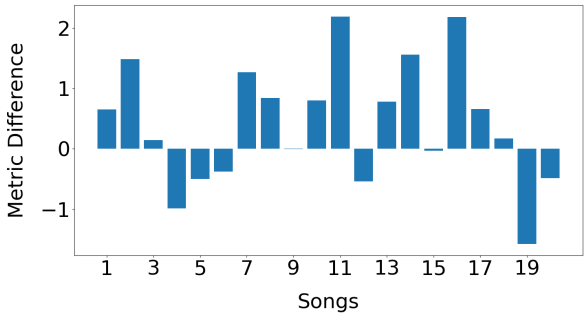

Figure 3. The result of subtracting the register histogram similarity of (C_{past}, C_{new}) from that of (C_{new}, C_{future})

Specifically, we choose 20 pieces of 12-bar musical segments from the AILabs-Pop1k7 dataset [2], which are MIDI transcriptions of polyphonic pop piano performances. The detailed settings of the VLI model is identical to our previous work [1]. For each segment, we use pre-trained VLI to insert 4-bar novel content at a manually-chosen boundary which is located at the edge of two *musical phrases*. MIDI files of all the original and expanded segments can be found at the demo website. [2]

3. EVALUATION

We use two metrics in our evaluation. We hypothesize that, if there is a sudden change in *rhythm* patterns at p going from C_{past} to C_{future}, such a sudden change should be preserved within the generated C_{new}. To quantify this, we use the *grooving pattern similarity* ($\mathcal{GS}$) proposed in [4] to compute the similarity in rhythm between $\{C_{\text{past}}, C_{\text{new}}\}$ (denoted as $\mathcal{GS}_1$) and between $\{C_{\text{new}}, C_{\text{future}}\}$ ($\mathcal{GS}_2$), respectively, and then compare the scores by subtracting the former from the latter (i.e., $\mathcal{GS}_2 - \mathcal{GS}_1$). A positive subtraction result indicates that the rhythm pattern of C_{new} is closer to that of C_{future}, otherwise to C_{past}. Besides, when there is a boundary, we expect the absolute value of the subtraction result to be a large number, assuming that only $\mathcal{GS}_1$ or $\mathcal{GS}_2$ would take a large value, but not both.

Across boundaries, there will also be changes in *pitch* patterns due to, e.g., splitting a complete chord into arpeggios or shifting melody line to another octave or key. To reflect these, we introduce a new metric called *register histogram similarity* ($\mathcal{RHS}$). Given a sequence of notes, we count the number of notes in each octave and construct a 7-dimensional histogram $\vec{h}$ from `C1` to `C7`. Given the histograms $\vec{h}_1, \vec{h}_2$ of two segments (e.g., $\{C_{\text{new}}, C_{\text{future}}\}$), we evaluate their similarity by their negative cross entropy (which lies in $(-\infty, 0]$, the closer to zero the more similar):

$$\mathcal{RHS}(\vec{h}_1, \vec{h}_2) = \sum_{i=0}^{7} h_{1,i} \log_2(h_{2,i}) \,. \qquad (1)$$

We similarly quantify the degree of changes in pitch by computing the $\mathcal{RHS}$ between $\{C_{\text{past}}, C_{\text{new}}\}$ and between $\{C_{\text{new}}, C_{\text{future}}\}$, respectively, and then subtracting them.

[2] https://tanchihpin0517.github.io/variable-length-piano-expansion/

Figures 2 and 3 show the result in $\mathcal{GS}$ and $\mathcal{RHS}$, respectively. We observe that the VLI model tends to keep the boundary in two ways: i) making C_{new} an extended content of C_{past} or C_{future} by imitating one of them, or ii) making C_{new} an independent segment which plays the role of a bridge connecting C_{past} and C_{future}. The boundary is often preserved well if the absolute subtraction result in $\mathcal{GS}$ or $\mathcal{RHS}$ is large, but not vice versa—informal listening shows that VLI is able to preserve the boundary even when the absolute subtraction results in both $\mathcal{GS}$ and $\mathcal{RHS}$ are low. We also find that VLI tends to make C_{new} more similar to C_{future} than to C_{past}, possibly because it is harder to extend a segment which has already a proper sense of ending, which is often the case for C_{past} in pop music.

4. CONCLUSION

In this paper, we have shown that the VLI model performs well for music score expansion in certain conditions. However, we currently use human-annotated boundaries for generation tasks. Selecting a target place for expanding is still an ongoing research topic. We aim to integrate music-structure analysis into the model and make it capable to decide where to insert content. Moreover, the metrics in our work are rather simplistic, describing the properties of boundaries fairly roughly. Finding other proper metrics is also one of our next steps.

5. REFERENCES

[1] C.-J. Chang *et al.*, "Variable-length music score infilling via XLNet and musically specialized positional encoding," in *Proc. ISMIR*, 2021.

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

[3] F. Lerdahl and R. S. Jackendoff, *A Generative Theory of Tonal Music*. MIT Press, 1996.

[4] S.-L. Wu and Y.-H. Yang, "The Jazz Transformer on the front line: Exploring the shortcomings of AI-composed music through quantitative measures," in *Proc. ISMIR*, 2020.