Audio Engineering Society

Late Breaking Demo Paper

Presented at the AES International Conference on
Artificial Intelligence and Machine Learning for Audio
2025 September 8–10, London, UK

Beat-Based Rhythm Quantization of MIDI Performances

Maximilian Wachter[1], Sebastian Murgul[1,2], and Michael Heizmann[2]

[1]*Klangio GmbH, Karlsruhe, Germany*
[2]*Institute of Industrial Information Technology, Karlsruhe Institute of Technology, Karlsruhe, Germany*

Correspondence should be addressed to Maximilian Wachter (max.wachter@klang.io)

ABSTRACT

We propose a transformer-based rhythm quantization model that incorporates beat and downbeat information to quantize MIDI performances into metrically-aligned, human-readable scores. We propose a beat-based preprocessing method that transfers score and performance data into a unified token representation. We optimize our model architecture and data representation and train on piano and guitar performances. Our model exceeds state-of-the-art performance based on the MUSTER metric.

1 Introduction

This paper introduces a novel transformer-based approach for beat-based rhythm quantization of symbolic MIDI performances. Rhythm quantization aims to recover the intended notated rhythm from expressive performances, a fundamental task in music information retrieval and automatic music trancription. Our method uniquely integrates beat information and performance timing into a unified tokenized input representation, enabling the model to learn rhythmic structure by modeling timing deviations relative to an underlying metrical grid. Unlike most quantization models, our model is capable of leveraging metronome information, which entails the possibility of completely eliminating the uncertainty of beat estimations.

2 Methods

To support diverse musical contexts, we designed a flexible preprocessing pipeline that uses beat estimations or ground truth beats as a priori information and adapts to multiple time signatures without requiring explicit time signature tokens. This enables the model to generalize and successfully quantize rhythms in time signatures unseen during training. We propose a confusion-based evaluation metric tailored to beat-aligned quantization, incorporating both onset accuracy and note value correctness, which is better suited to evaluating timing on a score-level than frame-based metrics.

We used the T5 transformer for our model architecture, which we subsequently optimized in terms of efficiency and quality. For training our model we used the ASAP dataset [1], a piano dataset containing performance

MIDI files with accompanying scores as well as beat and downbeat annotations. We trained our model on *N* measure sequences, assuming a one-to-one correspondence. Due to the sometimes poor alignment between scores and performances we filtered the dataset based on how well input and target sequences matched. We subsequently optimized the model based on input sequence length, note ordering, and data augmentation techniques such as pitch transposition, note deletion, and duration noise in order to maximize musical note onset accuracy.

3 Experiments

Our optimized checkpoint achieves an onset F1 score of 97.3% on the ASAP dataset and a note value accuracy of 83.3%. Extending beyond piano, we adapted the model to guitar performances using the Leduc dataset, demonstrating that instrument-specific training yields improved quantization results. This highlights the importance of capturing instrument-dependent rhythmic interpretation characteristics for precise quantization.

Finally, we benchmarked our approach against state-of-the-art methods using the test and training splits defined in the ACPAS dataset for computing the MUSTER metric [2]. In Table 1 we compared the onset and offset error rates ε_{onset} and ε_{offset} to other state-of-the-art probabilistic and deep learning-based approaches. We showed that our model is able to surpass state-of-the-art models in ε_{onset} by using beat annotations, while getting surpassed in ε_{offset} only by [3].

Table 1: Comparison of onset-time (ε_{onset}) and offset-time (ε_{offset}) error rates using the MUSTER metric.

Method	ε_{onset}	ε_{offset}
Neural Beat Tracking [4]	68.28	54.11
End-to-End PM2S [3]	15.55	**23.84**
HMMs (J-Pop) [5]	25.02	29.21
HMMs (classical) [5]	22.58	29.84
Our Model	**12.30**	28.30

4 Discussion

Our results confirm that explicit beat information can yield significant improvements in onset quantization performance. However, offset accuracy remains challenging due to our model's current inability to detect 32nd notes. Compared to similar models, ours excels in scenarios where explicit beat information is available like in cases where performances are recorded to a metronome. If metronome information is not available the performance is however highly dependent on beat estimation quality.

5 Summary

Our findings suggest that transformer-based models with beat-aware tokenization offer a powerful framework for expressive rhythm quantization across instruments and meters. Future work will explore incorporating explicit time signature modeling, and expanding to irregular meters and note values, aiming for even greater generalization and musical fidelity.

References

[1] Foscarin, F., Mcleod, A., Rigaux, P., Jacquemard, F., and Sakai, M., "ASAP: a dataset of aligned scores and performances for piano transcription," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[2] Nakamura, E., Benetos, E., Yoshii, K., and Dixon, S., "Towards Complete Polyphonic Music Transcription: Integrating Multi-Pitch Detection and Rhythm Quantization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[3] Beyer, T. and Dai, A., "End-to-end Piano Performance-MIDI to Score Conversion with Transformers," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[4] Liu, L., Kong, Q., Morfi, G., Benetos, E., et al., "Performance MIDI-to-score conversion by neural beat tracking," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[5] Shibata, K., Nakamura, E., and Yoshii, K., "Non-local musical statistics as guides for audio-to-score piano transcription," *Information Sciences*, 566, pp. 262–280, 2021.