YQX+: Benchmarking Expressive Piano Rendering with Probabilistic Models

Jinwen Zhou
Centre for Digital Music
Queen Mary University of London
London, UK
jinwen.zhou@qmul.ac.uk

Yuncong Xie
Dept. of MPAP
Steinhardt, New York University
New York, NY, USA
yx2568@nyu.edu

Haochen Wang
Dept. of MPAP
Steinhardt, New York University
New York, NY, USA
hw2346@nyu.edu

Aidan Hogg
Centre for Digital Music
Queen Mary University of London
London, UK
a.hogg@qmul.ac.uk

Simon Dixon
Centre for Digital Music
Queen Mary University of London
London, UK
s.e.dixon@qmul.ac.uk

Huan Zhang
Centre for Digital Music
Queen Mary University of London
London, UK
huan.zhang@qmul.ac.uk

***Abstract*—We present YQX+, a system framework for probabilistic modeling of piano performance expression. YQX+ formalizes the task of score-conditioned expression prediction using a multi-scale contextual feature representation, predicting note-level expressive parameters as beat period, velocity, timing deviation and articulation ratio. The YQX+ is extended from original YQX's Gaussian mixture modeling to XGBoost, Variational Autoencoders and flow-matching-based probabilistic rendering models for expressive performance.**

I. Introduction

Computational modelling of performance expression is a central challenge in music information retrieval (MIR). While recent systems have adopted deep generative models and multimodal architectures, much of this work has focused on performance generation, rather than predictive modelling of expressive variation under uncertainty.

A predecessor to this research direction is the YQX system [1], which won the 2008 Automatic Performance Rendering Contest (RenCon) [2]. YQX demonstrated that probabilistic modelling could capture expressive parameters such as tempo and loudness from score information. However, the system primarily produced deterministic predictions and was not designed for extensibility within modern machine learning frameworks.

We present **YQX+**, a reimplementation and extension of the YQX system, designed for expressive performance rendering in a contemporary machine learning setting. Unlike previous work that focus exclusively on rendering audio or symbolic sequences, YQX+ emphasises probabilistic modelling, enabling expressive variability to be represented and explored more systematically.

The improvements can be summarised as follows:

- A reimplementation of the original YQX with support for modern probabilistic frameworks and extended feature sets.
- Integration of a diverse set of models, from interpretable statistical approaches to state-of-the-art generative models, for rendering expressive piano performances.
- A flexible representation framework that supports systematic exploration of how different levels of score context contribute to expressive performance.

II. Methodology

A. Representation Framework

YQX+ remains the Feature-Modelling-Target framework, but extends both input feature set and the range of models: input features are extracted from symbolic scores (MusicXML), models learn mappings from features to expressive deviations, and targets specify how these deviations are rendered in performance.

1) Features: Following the original YQX system and inspired by midihum [3], we extended the original YQX feature set into four main categories. The subset of prominent features are shown in Figure 1.

- Pitch: Besides the basic pitch class and octave information, we extracted local intervals, contour direction, and relative position within the pitch range of a voice.
- Harmony: Harmony features include onset density, harmonic intervals to highest or lowest notes, and texture descriptors.
- Rhythm: Following the YQX system, we extracted local duration patterns, beat position within the measure, and normalized duration ratios within voice and piece.
- Phrase: Following Narmour's Implication-Realization theory [4] and implementation in YQX, we extracted phrase position, phrase length, and implication–realization (IR) features capturing melodic expectations.

These features incorporate both local context and longer structural information, enabling models to access hierarchical score cues relevant to expressive interpretation.

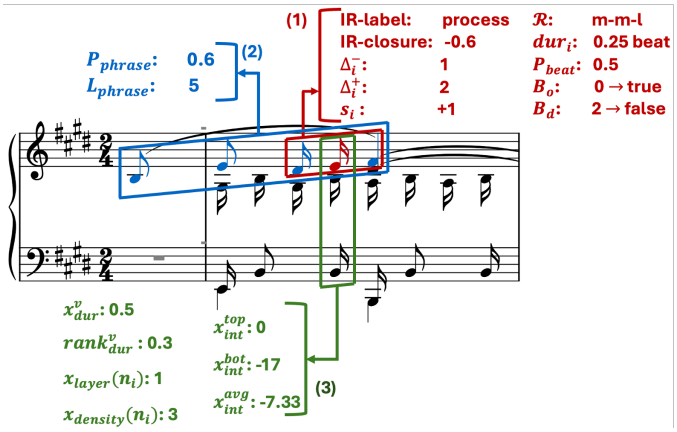

Fig. 1. An example showing some prominent features of: (1,red) the current note and the two adjacent notes; (2,blue) notes that belong to the same phrase as the current note; (3,green) notes simultaneous with current note.

2) Targets: Expressive performance is rendered using four note-level parameters inspired by the Basis Mixer framework [5]: beat period, velocity, timing and articulation ratio.

B. Probabilistic Modelling

To generate expressive performances, YQX+ integrates a range of modelling approaches, each with different strengths.

1) Gaussian Mixture Models (GMM): We replicate the original YQX design [1] using GPU-accelerated GMMs from the `gmm-gpu` library [6], which serve as a lightweight and interpretable baseline for probabilistic rendering.

2) XGBoost: Following the midihum [3], XGBoost regressors are trained as another baseline, predicting expressive parameters from score features and offering efficiency and robustness across feature sets.

3) Variational Autoencoder: A conditional β-VAE with transformer-based encoder–decoder architecture models expressive deviations as samples from a latent distribution. The encoder transforms note-wise contextual features into a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, from which a latent variable z is drawn. The decoder uses both z and the original context features to reconstruct the expressive targets. This allows rendering multiple expressive variants of the same score while maintaining consistency with structural context.

4) Conditional Flow Matching (CFM): Another rendering approach, CFM [7], treats performance expression as a transport process between noise and expressive parameters. A transformer network learns conditional velocity fields $v_0(x, t, f)$ that map Gaussian noise into expressive deviations, enabling high-fidelity, probabilistic rendering of performance variation [8], where x denotes the performance parameters, $t \in [0, 1]$ is a continuous variable that interpolates between the noise prior and the real data distribution, and f corresponds to the musical features.

III. Datasets

For model training, we relied on the ASAP [9] and ATEPP [10] datasets, which provide score–performance alignments in the form of MusicXML scores paired with performance MIDI recordings, covering 1,815 works from the Western classical repertoire. We investigated three types of feature configurations in order to examine how varying amounts of musical context influence model performance.

- Short context: This setup uses only local features derived from the target note and its immediate neighbours. It represents the narrowest temporal scope, focusing on detailed note-level characteristics without reference to broader musical structures.
- Long context: This setup expands the scope to include phrase level structure and long term dependency. Features capture multi-step melodic intervals, interval directions across two- or three-note windows, rhythmic patterns spanning multiple notes, as well as statistical descriptors of pitch and duration distributions at both the voice and piece levels.
- Full context: This setup integrates all available features, combining the extended context from the midihum feature set with technical indicators, thereby providing a rich, multi-scale representation of musical structure.

IV. Post-processing

The midi files are predicted by the model. Users can optionally specify their preferred tempo; if no target tempo is given, the model will use the tempo defined in the score, or default to 120 BPM if neither is available. There is no other symbolic quantisation, editing, mastering or other human intervention involved for submission.

References

[1] G. Widmer, S. Flossmann, and M. Grachten, "YQX plays Chopin," *AI Magazine*, vol. 30, no. 3, pp. 35–48, 2009.

[2] M. Hashida, T. Nakra, H. Katayose, T. Murao, K. Hirata, K. Suzuki, T. Kitahara, *et al.*, "Rencon: Performance rendering contest for automated music systems," in *Proceedings of the 10th international conference on music perception and cognition (ICMPC), Sapporo, Japan*, 2008.

[3] E. Waldemarsson, "midihum: A set of scripts for converting midi to humdrum." https://github.com/erwald/midihum, 2021. Accessed: 2025-07-29.

[4] M. T. Pearce, "Expectation in melody: The influence of context and learning," *Music Perception*, vol. 23, no. 5, pp. 377–405, 2006.

[5] C. E. Cancino-Chacón, *Computational Modeling of Expressive Music Performance with Linear and Non-linear Basis Function Models*. PhD thesis, Johannes Kepler University Linz, 2018.

[6] gmm-gpu developers, "gmm-gpu: Gaussian mixture models on gpu." https://pypi.org/project/gmm-gpu/, 2025. Accessed: 2025-07-30.

[7] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[8] O. Tal, A. Ziv, I. Gat, F. Kreuk, and Y. Adi, "Joint Audio and Symbolic Conditioning for Temporally Controlled Text-to-Music Generation," in *Proceeding of the 25t International Society on Music Information Retrieval (ISMIR)*, jun 2024.

[9] F. Foscarin, A. Mcleod, P. Rigaux, F. Jacquemard, and M. Sakai, "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.

[10] H. Zhang, J. Tang, S. Rafee, S. Dixon, and G. Fazekas, "ATEPP: A Dataset of Automatically Transcribed Expressive Piano Performance," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (Bengaluru, India), 2022.