EXPRESSIVE PIANO PERFORMANCE RENDERING FROM UNPAIRED DATA

Lenny Renault, Rémi Mignot and Axel Roebel

UMR9912 - STMS
Ircam, Sorbonne Université, CNRS, Ministère de la Culture
Paris, France
lenny.renault@ircam.fr | remi.mignot@ircam.fr | axel.roebel@ircam.fr

ABSTRACT

Recent advances in data-driven expressive performance rendering have enabled automatic models to reproduce the characteristics and the variability of human performances of musical compositions. However, these models need to be trained with aligned pairs of scores and performances and they rely notably on score-specific markings, which limits their scope of application. This work tackles the piano performance rendering task in a low-informed setting by only considering the score note information and without aligned data. The proposed model relies on an adversarial training where the basic score notes properties are modified in order to reproduce the expressive qualities contained in a dataset of real performances. First results for unaligned score-to-performance rendering are presented through a conducted listening test. While the interpretation quality is not on par with highly-supervised methods and human renditions, our method shows promising results for transferring realistic expressivity into scores.

1. INTRODUCTION

Performance rendering is the task of imbuing a music score with expressive features as if a musician performed the score in a way to bring out emotional qualities. To get an expressive rendition of the music, performers have the liberty to shape sound parameters that are not explicitly described by the written score [1]: for piano pieces, musicians make an interpretation of the score by mainly reshaping the timing, articulation and nuance of the notes. An automated system that can reproduce such a complex and artistic behavior can find its usage in assisting composers for obtaining musical renditions of their pieces.

Previous works for the task used data-driven methods to predict performance features that enhance the score note indications [2, 3, 1, 4]. More recently, *Variational Auto-Encoders* (VAE) conditioned on score features have proven to be successful at modeling the diversity in performance expressivity, as several renditions of the same piece are conceivable [5, 6, 7, 8]. The performance features are defined as the difference in timing, articulation, and velocity of the played notes compared to the exact rendition of the score [9]. However, obtaining such features requires the collection of *Musical Instrument Digital Interface* (MIDI) performances with their associated digital scores and to align them at note-level [10, 11]. These required matching and alignment steps limit the amount of data available for training [12] and the application of the models to piano music, where performance MIDI data can be collected more easily [13]. Also, most of these works are highly-informed as they take different markings in the digital scores into account for guiding the expressive rendering, such as rests, beat information, hand part, position in the measure, key and time signatures, articulation and ornament markings, slurs or beams. This reliance on markings specific to the sheet music format hinders the usage of these models in modern music production frameworks (DAW, sequencers) where MIDI data are directly manipulated without using such markings.

Concurrently, *Generative Adversarial Networks* (GAN) have been successfully applied for various tasks transferring data from one domain to another without aligned pairs, such as image-to-image translation [14], audio timbre matching [15] or music genre transfer [16]. In the light of such results, this work attempts to address expressive performance rendering as a domain transfer task, by transforming MIDI scores into human-like performances without supervision on the performance features and reliance on score markings. To this end, an adversarial approach is employed to map the outputs of a low-informed performance rendering model to the distribution of human performances, without providing matching pairs of scores and performances. Trained on publicly available datasets, the proposed method and its experiments are presented here, including an early subjective evaluation.

The experiments show promising results for the method as it can infer expressive qualities into scores, although not with the same amount of naturalness as in performances rendered by real pianists and by a highly-informed supervised baseline. Accompanying this paper, audio samples are provided online [1].

2. PROPOSED APPROACH

The proposed approach, illustrated in Figure 1, is composed of a performance rendering model G that takes a score X as input and produces an expressive interpretation $\tilde{X}$. The rendered performances are fed into a discriminator D, among performances Y from a dataset of recorded human performances. The performance rendering model and the discriminator have opposed objectives, as the discriminator D aims to differentiate the real performances from the ones rendered by the model G, while the latter tries to produce performances indistinguishable from the real ones.

2.1. Data formatting

Both the scores X and real performances Y are encoded as sequences of N notes with the minimal amount of features needed for describing them:

$$\mathbf{X} = \{x_n\}_{n\leq N} = \{p_n, o_n, d_n, v_n\}_{n\leq N}. \tag{1}$$

[1] http://renault.gitlab-pages.ircam.fr/dafx23

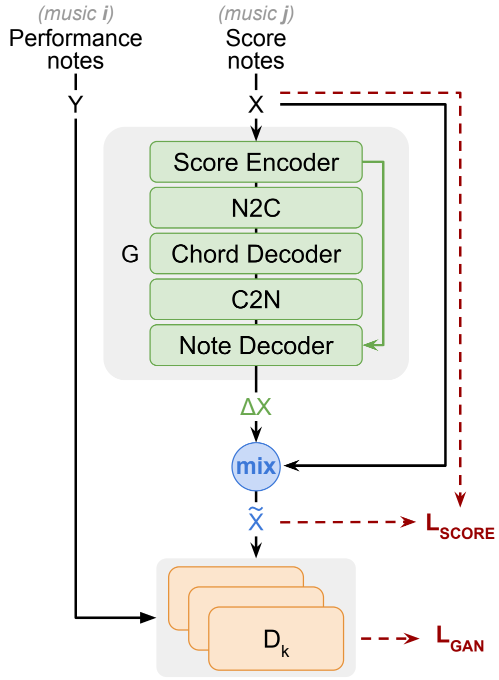

Figure 1: *Training pipeline for the proposed model: the mix function modifies the score X with features output by the performance rendering model G (in green), in order to deceive the discriminators D_k (in orange). During training, the unaligned score X and performance Y are drawn at random from their respective sets.*

The notes are ordered by their absolute onset time: for the n-th note, p_n is its normalized MIDI pitch, o_n its delta-time with the previous note onset, or relative *inter-onset-interval* (IOI), capped at 4s, d_n its duration in absolute time and v_n its normalized MIDI velocity.

2.2. Performance rendering model

The performance rendering model G predicts modifying features $\Delta X = G(X)$ from the score note features in order to modify them into performance-like note features $\tilde{X}$ through the mix function:

$$\begin{aligned}\tilde{X} &= \text{mix}(X, G(X)) \\ &= \{p_n, o_n + \delta o_n, d_n \times \delta d_n, v_n \times \delta v_n\}_{n \leq N},\end{aligned} \tag{2}$$

with δo_n the micro-onset timing, δd_n the articulation and δv_n the expressive velocity of the n-th score note.

These modifying features are obtained by first processing the note-wise score features with a convolutional Score Encoder. Then, the same hierarchical modeling from [5] is applied: the note-wise features are merged into chord-wise features, which enables a more coherent modeling of the full sequence. This note-to-chord operation, or $N2C$, is performed by average pooling the features of simultaneous notes into a common chord-wise feature. The inverse operation $C2N$ can later convert chord-wise features into note-wise features by duplicating the chord feature for each of its notes. On the contrary of hierarchical strategies employed in other works [7, 8], the note-to-chord alignment matrix required for $N2C$ and $C2N$ can be directly extracted from our low-informed MIDI data representation, using the sequence of relative IOI $\{o_n\}_{n \leq N}$. Further implementation details on the $N2C$ and $C2N$ operations can be found in [5].

Before returning to the note-granularity, the chord-wise features are further processed by a Chord Decoder, which is a *Convolutional Recurrent Neural Network* (CRNN) with a bidirectional *Gated Recurrent Unit* (GRU) layer. Finally, fine-grained adjustments at note-level are made with the Note Decoder and a skip connection from the note-wise score encoding. The final micro-onset timing δo_n is obtained through a linear activation function, while the articulation δd_n and the expressive velocity δv_n are mapped to $[0.25, 4]$ with a scaled sigmoid function.

2.3. Discriminator

Taking inspiration from speech processing using discriminators with a multi-scale architecture [17], we use $k = 3$ discriminators D_k with identical architectures, mirrored from the performance rendering model, with the exceptions of the $N2C$ and $C2N$ operations, as chords in real performances are not as easily defined as in scores. Each discriminator is fed with a downsampled sequence of (real or rendered) performance notes by average pooling with sizes $\{1, 3, 9\}$. Discriminators with longer pool sizes look at features at higher levels in the performances and thus, can help transferring such knowledge and long-term coherence to the performance rendering model G. To stabilize the GAN training, gaussian noise is added to the inputs of the discriminators, as in [16].

2.4. Loss functions

The least-square variant of the GAN objective (LSGAN) is used to train the discriminators and the performance rendering model. Their respective loss functions L_{D_k} and $L_{G,gan}$ are defined as:

$$\begin{aligned}L_{D_k} &= \mathop{\mathbb{E}}_{Y \sim p_{perf}} [\|D_k(Y) - 1\|_2] + \mathop{\mathbb{E}}_{X \sim p_{score}} [\|D_k(G(X))\|_2], \\ L_{G,gan} &= \mathop{\mathbb{E}}_{X \sim p_{score}} \left[\sum_{k=1,2,3} \|D_k(G(X)) - 1\|_2 \right].\end{aligned} \tag{3}$$

We have observed that the instability of the vanilla adversarial training may lead the performance rendering model to displace the notes in extreme values, causing the original piece to be unrecognizable. To ensure that the performances remain fairly close to their scores, an additional regularization term L_{score} is added:

$$L_{score}(X) = \boldsymbol{\lambda}_{score} \left\| \frac{G(X) - X}{X} \right\|_2, \tag{4}$$

with $\boldsymbol{\lambda}_{score}$ a fixed vector weighting how much each performance component (timing, articulation, velocity) can deviate from the score indication. Here, $\boldsymbol{\lambda}_{score} = \{1, 1, 0.1\}$.

The total loss for the performance rendering model G is:

$$L_G(X) = \lambda_{gan} L_{G,gan}(X) + L_{score}(X), \tag{5}$$

with λ_{gan} the balance between the GAN objective and the score regularization loss. This balance is decisive for the final behavior of G since the two loss components have opposite influences on its training: L_{score} refrains G from modifying the scores while $L_{G,gan}$ encourages exploring different interpretations in order to deceive the discriminator. In our experiments, $\lambda_{gan} = 2$.

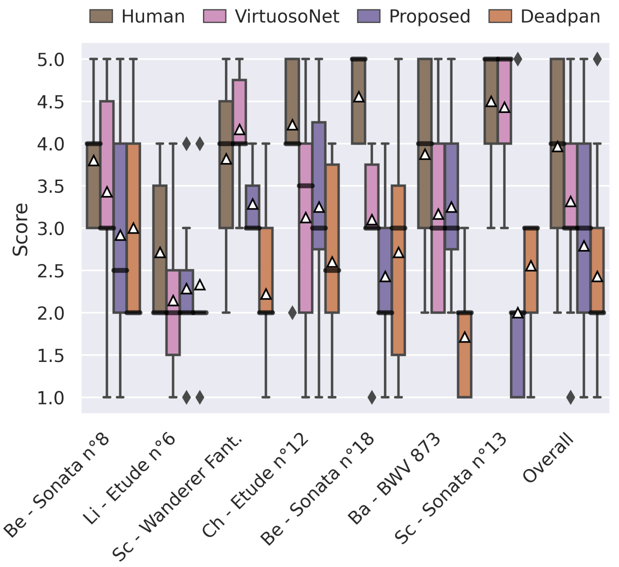

Figure 2: *Box-plot of the* Mean Opinion Score *(MOS) of the different performance rendering methods (overall and piece-wise). The thickened bars indicate the median values while the white triangles indicates the mean values. The composers are Bach (Ba), Beethoven (Be), Chopin (Ch), Liszt (Li) and Schubert (Sc).*

3. EXPERIMENTS

3.1. Score and performance datasets

The proposed approach was trained and evaluated using the scores from the ASAP dataset [12] and all performances from the MAESTRO dataset (v3.0.0) [13], which are both publicly available. The human performances from MAESTRO were recorded in MIDI format using Yamaha Disklaviers. The ASAP dataset has notably matched a set of these performances with their original scores at note-level, and has thus been used to some extent in previous performance rendering works [5]. However, since the proposed method does not require aligned scores and performance, the entirety of both datasets can be used, which amounts for 962 training performances, 137 validation performances, 107 training scores, 15 validation scores and 35 test scores (following the train-validation-test split of [18]).

The velocity indications were kept from the ASAP scores in MIDI format, which can either be constant throughout the piece or mapped from the score nuances and markings using simple rules. The scores and performances are split into segments of 128 consecutive notes, with random pitch shifting during training by ±7 semi-tones, as in [6]. Validation data is used to monitor and avoid potential over-fitting of the performance rendering model by reproducing the training performances from their corresponding scores.

3.2. Early subjective evaluation

A listening test has been conducted to evaluate the interpretation quality of the performances rendered by our model. 7 scores from the ASAP test subset were selected, covering 5 different composers. 4 performances were generated by different methods for each score: a corresponding human performance from the MAESTRO dataset (**Human**), the direct export of the MIDI score (**Deadpan**), a rendition by our approach (**Proposed**) and a rendition from the graph-based variant of **VirtuosoNet** [8], a highly-informed model using score markings in MusicXML format and is trained with an private dataset of 226 scores matched and aligned with MAESTRO performances, which is larger than ASAP. The first 20s of each performance were synthesized using the Arturia Piano V3 software [2], a physical-based piano synthesizer. 19 professional audio and piano players were asked to rate the naturalness of the presented performances, using a 5-point Likert scale (from 1 - Bad, to 5 - Excellent). Each trial randomly presented 3 different performances from each method. Results are reported in Figure 2.

The Holm-Bonferroni corrected two-sided Mann-Whitney U tests indicate a statistical difference at $\alpha = 0.05$ between the Human rendition and each other methods, and between VirtuosoNet and Deadpan. The overall results show that the proposed approach does enhance the scores with expressive features in comparison to the raw rendition of the piece, but still not with the same amount of naturalness as actual pianists and the highly-informed VirtuosoNet. This was to be expected as our proposed unsupervised training task without score markings is harder than the training objectives of VirtuosoNet, for about the same quantity of training data. By examining the ratings piece-wise, one can notice the poorer renditions of the proposed method for slower tracks (Schubert's 13th Sonata and Beethoven's 18th Sonata). This may suggest that the model has a mode collapse on faster paced music and that it applies similar modifying features on every tracks, which renders inappropriate performances for slower musical pieces.

4. FUTURE WORK

As suggested by the subjective evaluation, the model lacks in understanding the musical content of a score and can apply inappropriate performance features. Also, on the contrary of the most recent supervised performance rendering methods [8, 5, 6], the model does not allow for external controls (tempo, articulation, nuance) on the rendering process. Both issues can be addressed by organizing the performances into sub-domains with either domain labels (such as composer or genre) or with extracted performance features (note density, statistics on durations and velocities)[19].

Moreover, the present work only focuses on classical piano music to be comparable with previous supervised approaches, but without reliance on training pairs, the approach can be extended to render symbolic performances for other genres and instruments.

Finally, GAN enable unsupervised cross-modal domain transfer where the target domain can be in a different modality from the source domain. By including a differentiable sound synthesizer [20] after the performance rendering model and using a audio-based performance discriminator, the model could potentially render scores with expressive features by learning from performances in the audio domain instead of MIDI.

5. CONCLUSIONS

This work presents a performance rendering model for converting piano scores into expressive performances, without supervised training on performance features nor relying on sheet music markings. Using a performance discriminator, the model reshapes the

[2] https://www.arturia.com/products/software-instruments/piano-v/overview

basic score note properties into a sequence of expressive notes. Trained on publicly available datasets of scores and performances, the approach shows expressive qualities in the performance renditions compared to the plain score, although not with the same quality as a fully supervised approach, according to a conducted listening test. Still, by removing the reliance on training with paired data and on score markings, the approach can be further used in broader settings with music in different modalities and genres.

6. ACKNOWLEDGMENTS

This work was supported by European Union's Horizon 2020 research and innovation programme under grant number 951911 - AI4Media.

7. REFERENCES

[1] Carlos E Cancino-Chacón, Maarten Grachten, Werner Goebl, and Gerhard Widmer, "Computational models of expressive music performance: A comprehensive and critical review," *Frontiers in Digital Humanities*, vol. 5, 2018.

[2] Iman Malik and Carl Henrik Ek, "Neural translation of musical style," in *Workshop on Machine Learning for Creativity and Design, Neural Information Processing Systems (NIPS)*, Long Beach, California, USA, Dec. 8, 2017.

[3] Fábio José Muneratti Ortega et al., *A machine learning approach to computer modeling of musical expression for performance learning and practice*, Ph.D. thesis, Universitat Pompeu Fabra, 2022.

[4] Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep performer: Score-to-audio music performance synthesis," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 951–955.

[5] Seungyeon Rhyu, Sarah Kim, and Kyogu Lee, "Sketching the expression: Flexible rendering of expressive piano performance with self-supervised learning," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, Dec. 4-8, 2022.

[6] Akira Maezawa, Kazuhiko Yamamoto, and Takuya Fujishima, "Rendering music performance with interpretation variations using conditional variational RNN," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, Nov. 4-8, 2019, pp. 855–861.

[7] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 908–915.

[8] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, and Juhan Nam, "Graph neural network for music score data and modeling expressive piano performance," in *Proc. of the International Conference on Machine Learning (ICML)*, Long Beach, California, USA, June. 9-15, 2019, PMLR, pp. 3060–3070.

[9] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, and Juhan Nam, "Score and performance features for rendering expressive music performances," in *Proc. of the Music Encoding Conference*, Vienna, Austria, May 2019.

[10] Francesco Foscarin, Emmanouil Karystinaios, Silvan David Peter, Carlos Cancino-Chacón, Maarten Grachten, and Gerhard Widmer, "The match file format: Encoding alignments between scores and performances," in *Proc. of the Music Encoding Conference*, Halifax, Canada, May. 19-22, 2022.

[11] Eita Nakamura, Kazuyoshi Yoshii, and Haruhiro Katayose, "Performance error detection and post-processing for fast and accurate symbolic music alignment," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, Oct. 2017, pp. 347–353.

[12] Francesco Foscarin, Andrew Mcleod, Philippe Rigaux, Florent Jacquemard, and Masahiko Sakai, "ASAP: a dataset of aligned scores and performances for piano transcription," in *Proc. of the International Society for Music Information Retrieval (ISMIR)*, Montreal / Virtual, Canada, Oct. 2020.

[13] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, May 2019, OpenReview.net.

[14] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022.

[15] Alec Wright, Vesa Välimäki, and Lauri Juvela, "Adversarial guitar amplifier modelling with unpaired data," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, June. 4-10, 2023, IEEE.

[16] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Sumu Zhao, "Symbolic music genre transfer with CycleGAN," in *International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2018, pp. 786–793.

[17] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems (NIPS)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[18] Lele Liu, Veronica Morfi, and Emmanouil Benetos, "ACPAS: a dataset of aligned classical piano audio and scores for audio-to-score transcription," in *Late-Breaking Demos of the International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[19] Kristy Choi, Curtis Hawthorne, Ian Simon, Monica Dinculescu, and Jesse Engel, "Encoding musical style with transformer autoencoders," in *Proc. of the International Conference on Machine Learning (ICML)*, Jul. 13–18, 2020, pp. 1899–1908.

[20] Lenny Renault, Rémi Mignot, and Axel Roebel, "Differentiable piano model for midi-to-audio performance synthesis," in *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Vienna, Austria, Sep. 2022, pp. 232–239.