SKETCHING THE EXPRESSION: FLEXIBLE RENDERING OF EXPRESSIVE PIANO PERFORMANCE WITH SELF-SUPERVISED LEARNING

Seungyeon Rhyu[1] **Sarah Kim**[2] **Kyogu Lee**[1,3]
[1] Music and Audio Research Group (MARG), Seoul National University, South Korea
[2] Krust Universe, South Korea
[3] Graduate School of AI, AI Institute, Seoul National University, South Korea
rsy1026@snu.ac.kr, estelle.kim@krustuniverse.com, kglee@snu.ac.kr

ABSTRACT

We propose a system for rendering a symbolic piano performance with flexible musical expression. It is necessary to actively control musical expression for creating a new music performance that conveys various emotions or nuances. However, previous approaches were limited to following the composer's guidelines of musical expression or dealing with only a part of the musical attributes. We aim to disentangle the entire musical expression and structural attribute of piano performance using a conditional VAE framework. It stochastically generates expressive parameters from latent representations and given note structures. In addition, we employ self-supervised approaches that force the latent variables to represent target attributes. Finally, we leverage a two-step encoder and decoder that learn hierarchical dependency to enhance the naturalness of the output. Experimental results show that our system can stably generate performance parameters relevant to the given musical scores, learn disentangled representations, and control musical attributes independently of each other.

1. INTRODUCTION

Computational modeling of expressive music performance focuses on mimicking human behaviors that convey the music [1, 2]. For piano performance, one common task is to *render* an expressive performance from a quantized musical score. It aims to reproduce the loudness and timing of musical notes that fits to the given score. Most of the conventional studies have used musical scores of Western piano music that includes sufficient amount of guidelines for musical expressions [3–6]. Recent studies using deep learning methods have successfully rendered plausible piano performances that are comparable to those of professional pianists from the given Classical scores [7–9].

More recently, it has increased attention to *controlling* music performance by manipulating one or more *disentangled* representations from a generative model. These representations are sensitive to the variation of certain factors while invariant to other factors [10]. Maezawa *et al.* aimed to control a performer's interpretation through a conditional variational recurrent neural network (CVRNN) [11]. They intended to disentangle a time-variant representation of the personal interpretation. In the acoustic domain, Tan *et al.* proposed a generative model based on a Gaussian mixture variational autoencoder (GM-VAE) that separately controlled dynamics and articulations of the notes [12]. Their novelty lied in learning multiple representations of high-level attributes from the low-level spectrogram.

However, these studies have constrained musical creativity. Maezawa *et al.* controlled musical expression only through quantized features from the musical scores. Tan *et al.* did not consider controlling tempo or timing with a latent representation. These methods may have restricted any potential for rendering piano performances with flexible musical expression. Musical creativity can be expanded not only by composers but also by performers who can elastically choose various strategies to highlight multiple nuances or emotions [13–15]. Moreover, the music generation field can be also broadened if static music created by automatic composition systems can be easily colored with realistic and elastic expression [16].

Therefore, we attempt a new approach that renders piano performances with flexible musical expressions. We disregard a typical assumption from previous studies that a performer must follow a composer's intent [4, 17–19]. According to the literature, performers learn to identify or imitate "expressive models", or *explicit planning*, of existing piano performances [20]. We focus on this attribute, defining it as a higher-level *sketch* of the expressive attributes (i.e. dynamics, articulation, and tempo [21]) that the performer draws based on a personal interpretation of the musical piece [4, 11, 20]. We also assume that the remaining attribute represents common performing strategies that are connected to certain musical patterns, while these strategies slightly differ across performers [22, 23]. We call this attribute as a *structural attribute* that belongs to given note structures of a musical piece.

In this study, we propose a generative model that can

flexibly control the entire musical expression, or the explicit planning, of symbolic piano performance [1]. Our system is based on a conditional variational autoencoder (CVAE) that is modified for sequential data [11, 24]. The system generates multiple parameters of piano performance from a note structure of a musical passage, using disentangled representations for the explicit planning and structural attribute.

We employ a self-supervised learning framework to force the latent representations to learn our target attributes [24–26]. In addition, we facilitate independent control of the three expressive attributes–dynamics, articulation, and tempo–by utilizing an existing method that aligns the latent code with a target attribute [27,28]. Finally, we design a novel mechanism that intuitively models a polyphonic structure of piano performance. In particular, we insert intermediate steps for *chordwise* encoding and decoding of the piano performance to our encoder-decoder architecture, where a *chord* denotes a group of simultaneous notes.

Our approach has several contributions as follows: 1) Our system aims to control musical expression while maintaining any characteristics induced by a given musical structure; 2) We use self-supervised learning where new supervisory signals are involved in regularizing the latent representations effectively; 3) Our system aims to control multiple expressive attributes independently of each other; 4) Lastly, we leverage an intermediate step that projects a notewise representation into the chordwise in the middle of our system to intuitively model the polyphonic structure of piano performance.

2. PROPOSED METHODS

We aim to build a generative model that factorizes expressive piano performance as the explicit planning and structural attribute. The model is based on a conditional variational autoencoder (CVAE) that reproduces performance parameters based on a given musical structure.

2.1 Data Representation

We extract features that represent a human performance and the corresponding musical score, following the conventional studies [11, 19, 29].

Performance Features. We extract three features that represent the expressive attributes of each performed note, respectively: **MIDIVelocity** is a MIDI velocity value that ranges from 24 to 104. **IOIRatio** represents an instantaneous variation in tempo. We compute an inter-onset-interval (IOI) between the onset of a note and the mean onset of the *previous* chord for both a performed note and the corresponding score note. Then, a ratio of performed IOI to score IOI is calculated, clipped between 0.125 and 8, and converted into a logarithmic scale [4]. **Articulation** represents how much a note is shortened or lengthened compared to the instantaneous tempo. It is a ratio of a performed duration to an IOI value between the onset of a note and mean onset of the *next* chord [19]. It is clipped between 0.25 and 4 and converted into a logarithmic scale.

Score Features. The features for a musical score represent eight categorical attributes for how the notes are composed: **Pitch** is a MIDI index number that ranges from 21 to 108. **RelDuration** and **RelIOI** are 11-class attributes of a quantized duration and IOI between a note onset and a previous chord, respectively. They range from 1 to 11, and each class represents a multiple of a 16th note's length with respect to a given tempo [30, 31]. **IsTopVoice** is a binary attribute of whether the note is the uppermost voice. It is heuristically computed regarding pitches and durations of surrounding notes. **PositionInChord** and **NumInChord** are 11-class attributes of a positional index of a note within its chord and the total number of notes in that chord, respectively, that range from 1 to 11. An index 1 for PositionInChord denotes the most bottom position. **Staff** is a binary attribute of the staff of a note, either of the G clef or F clef. **IsDownbeat** is a binary attribute of whether a note is at a downbeat or not.

2.2 Modeling Musical Hierarchy

Inspired by previous studies [4, 8, 9, 32], we build a two-step encoder and decoder: An encoder models both notewise and chordwise dependencies of the inputs, and a decoder reconstructs the notewise dependency from the chordwise representation and the notewise condition. We denote a *chord* as a group of notes that are hit simultaneously, regardless of the staff, so that they sound together at an instant time [33]. Thus, learning the chordwise dependency is analogous to direct modeling of the temporal progression of the piano performance. Let $\mathcal{M} \in \mathbb{R}^{C\times N}$ be a matrix that aligns serialized notes to their polyphonic structure, where C and N are the number of chords and the number of notes, respectively. Within the encoder, the notewise representation is sequentially average-pooled by $\mathcal{M}$ with dynamic kernel sizes where each size represents the number of notes in each chord. We denote this operation as *N2C*. In this way, we can directly model chord-level dependency of the note-level expressive parameters [32]. In contrast, the decoder extends the chordwise representation from the encoder back to the notewise using the transposed alignment matrix $\mathcal{M}^T$, of which process we denote as *C2N*. Along this, the notewise embedding of the score features replenishes the notewise information for the output. Consequently, notes in the same chord *share* any information of their corresponding chord, while maintaining their differences by the conditional score features:

$$\mathrm{N2C}(e) = \frac{\mathcal{M}\cdot e}{\sum_{n=1}^{N}\mathcal{M}_{n,1:C}}, \quad \mathrm{C2N}(e) = \mathcal{M}^{\mathrm{T}}\cdot e \qquad (1)$$

where e denotes a notewise or chordwise representation.

2.3 Overall Network Architecture

Our proposed network is generally based on the conditional VAE framework [34, 35]. Concretely, we use the sequential VAE that is modified for generation of sequential data [11,24,36]. Let $x = \{x_n\}_{n=1}^{N}$ be a sequence of the

[1] https://github.com/rsy1026/sketching_piano_expression

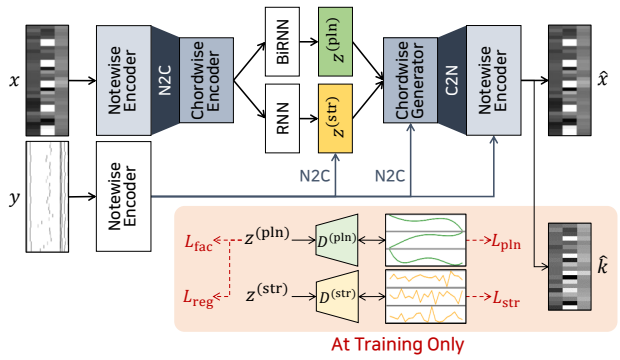


Figure 1: Overall architecture of the proposed system. The orange box includes the auxiliary tasks only for training.

performance features, and $y = \{y_n\}_{n=1}^{N}$ be a sequence of the conditional score features. Our network has two *chord-wise* latent variables $z^{(\text{pln})} = \{z_c^{(\text{pln})}\}_{c=1}^{C} \in \mathbb{R}^{C \times d^{(\text{pln})}}$ and $z^{(\text{str})} = \{z_c^{(\text{str})}\}_{c=1}^{C} \in \mathbb{R}^{C \times d^{(\text{str})}}$ that represent explicit planning and structural attribute, where $d^{(\text{pln})}$ and $d^{(\text{str})}$ are the sizes of $z^{(\text{pln})}$ and $z^{(\text{str})}$, respectively. Our network generates notewise performance parameters x from these latent variables and given score features y. The overall architecture of our proposed system is illustrated in Figure 1.

Generation. A probabilistic generator parameterized by θ produces the note-level performance parameters x from the two latent variables $z^{(\text{pln})}$ and $z^{(\text{str})}$ with the given condition y. We note that the latent variables are in chord-level. This decreases a computational cost and also enables intuitive modeling of polyphonic piano performance where each time step represents a stack of notes and the simultaneous notes share common characteristics [8]:

$$p_\theta(x, y, z^{(\text{pln})}, z^{(\text{str})}) = p_\theta(x|z^{(\text{pln})}, z^{(\text{str})}, y) \\ p_\theta(z^{(\text{pln})}) \prod_{c=1}^{C} p_\theta(z_c^{(\text{str})}|z_{<c}^{(\text{str})}, y_{\leq c}^{(\text{chd})}) \quad (2)$$

where $y^{(\text{chd})} = \text{N2C}(e_y)$ is the chordwise embedding, and e_y is the notewise embedding for y. We assume that the prior of $z_c^{(\text{pln})}$ is a standard normal distribution. In contrast, $z_c^{(\text{str})}$ is sampled from a sequential prior [24, 36, 37], conditioned on both previous latent variables and chordwise score features: $z_c^{(\text{str})} \sim \mathcal{N}(\mu^{(\text{prior})}, \text{diag}(\sigma^{(\text{prior})^2})$, where $[\mu^{(\text{prior})}, \sigma^{(\text{prior})}] = f^{(\text{prior})}(z_{<c}^{(\text{str})}, y_{\leq c}^{(\text{chd})})$, and $f^{(\text{prior})}$ is a unidirectional recurrent neural network. The latent representations and $y^{(\text{chd})}$ pass through the decoder as shown in Figure 1. During training, the model predicts the intermediate chordwise output that is computed as $\text{N2C}(x)$. This is to enhance reconstruction power of our system, propagating accurate information of chord-level attributes to the final decoder. The intermediate activation is then extended to the notewise through the C2N operation. The note-level parameters are generated autoregressively based on this activation and the notewise score feature. We use teacher forcing during training [38].

Inference. A probabilistic encoder parameterized by ϕ approximates the posterior distibutions of the latent representations $z^{(\text{pln})}$ and $z^{(\text{str})}$ from the performance input x and conditional score input y:

$$q_\phi(z^{(\text{pln})}, z^{(\text{str})}|x, y) = q_\phi(z^{(\text{pln})}|x^{(\text{chd})}) \\ \prod_{c=1}^{C} q_\phi(z_c^{(\text{str})}|x_{\leq c}^{(\text{chd})}, y_{\leq c}^{(\text{chd})}) \quad (3)$$

where $x^{(\text{chd})} = \text{N2C}(e_x)$ is the chordwise embedding, and e_x is the notewise embedding for x. The posterior distributions of $z_c^{(\text{pln})}$ and $z_c^{(\text{str})}$ are approximated by distribution parameters encoded by $f^{(\text{pln})}(x^{(\text{chd})})$ and $f^{(\text{str})}(x^{(\text{chd})}, y^{(\text{chd})})$, where $f^{(\text{pln})}$ and $f^{(\text{str})}$ are bidirectional and unidirectional recurrent neural networks, respectively. We note that $z^{(\text{pln})}$ is independent of the score features y. This allows a flexible transfer of the explicit planning among other musical pieces. On the other hand, $z^{(\text{str})}$ is constrained by y since the structural attributes are dependent on the note structure.

Training. We train the models p_θ and q_ϕ by approximating marginal distributions of the performance features x conditioned on the score features y. This requires to maximize negative evidence lower bound (ELBO) that includes regularization force by Kullback–Leibler divergence [34]:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z^{(\text{pln})}, z^{(\text{str})}|x,y)} \left[\log p_\theta(x|z^{(\text{pln})}, z^{(\text{str})}, y)\right] \\ + \mathbb{E}_{q_\phi(z^{(\text{pln})}, z^{(\text{str})}|x,y)} \left[\log p_\theta(k|z^{(\text{pln})}, z^{(\text{str})}, y)\right] \\ - \text{KL}(q_\phi(z^{(\text{pln})}|x) \| p_\theta(z^{(\text{pln})})) \\ - \sum_{c=1}^{C} \text{KL}(q_\phi(z_c^{(\text{str})}|x_{\leq c}^{(\text{chd})}, y_{\leq c}^{(\text{chd})}) \| p_\theta(z_c^{(\text{str})}|z_{<c}^{(\text{str})}, y_{\leq c}^{(\text{chd})})) \quad (4)$$

where $k = \text{N2C}(x)$ is the chordwise performance features.

2.4 Regularizing the Latent Variables

We enhance disentanglement of the latent representations $z^{(\text{pln})}$ and $z^{(\text{str})}$ using four regularization tasks [24].

Prediction Tasks. We extract new supervisory signals for additional prediction tasks from the input data [24]. We define a signal of explicit planning $I^{(\text{pln})}$ as a set of smoothed contours of the expressive parameters. It is extracted as a polynomial function predicted from the chordwise performance parameters k. We also derive a signal of structural attribute as $I^{(\text{str})} = \text{sign}(k - I^{(\text{pln})})$ which represents normalized directions of the performance parameters. We train two discriminators $D^{(\text{pln})}$ and $D^{(\text{str})}$ that directly receive $z^{(\text{pln})}$ and $z^{(\text{str})}$, respectively. $D^{(\text{pln})}$ is composed of A sub-discriminators where each discriminator $D_a^{(\text{pln})}$ predicts a signal $I_a^{(\text{pln})}$ for each expressive attribute a from $z_a^{(\text{pln})} \in \mathbb{R}^{C \times (d^{(\text{pln})}/A)}$, where $z_a^{(\text{pln})}$ is a constituent part of $z^{(\text{pln})}$, and A is the number of expressive attributes. This setting is for a clear disentanglement among the expressive attributes. On the other hand, $D^{(\text{str})}$ predicts the signal $I^{(\text{str})}$ at once for all expressive attributes that belong to the same musical structure. All discriminators are jointly trained with the generative model, and the costs $\mathcal{L}_{\text{pln}}$ and $\mathcal{L}_{\text{str}}$ are minimized

as $\mathcal{L}_{\text{pln}} = \frac{1}{A}\sum_a \text{MSE}(D_a^{(\text{pln})}(z_a^{(\text{pln})}), I_a^{(\text{pln})})$ and $\mathcal{L}_{\text{str}} = \text{MSE}(D^{(\text{str})}(z^{(\text{str})}), I^{(\text{str})})$, respectively.

Factorizing Latent Variables. We further constrain a generator to guarantee that $z^{(\text{pln})}$ delivers correct information regardless of $z^{(\text{str})}$ [39]. During training, we sample a new output $\tilde{x}$ using $z^{(\text{pln})} \sim q_\phi(z^{(\text{pln})}|x)$ and $\tilde{z}^{(\text{str})} \sim p_\theta(z^{(\text{str})})$. Then, we re-infer $\tilde{z}^{(\text{pln})} \sim q_\phi(\tilde{z}^{(\text{pln})}|\tilde{x})$ to estimate the superversory signal $I^{(\text{pln})}$. This prediction loss is backpropagated only through the generator:

$$\mathcal{L}_{\text{fac}} = \frac{1}{A}\sum_a \text{MSE}(D_a^{(\text{pln})}(\tilde{z}_a^{(\text{pln})}), I_a^{(\text{pln})}) \quad (5)$$

Aligning Latent Variables with Factors. Finally, we enable the "sliding-fader" control of the expressive attributes [28]. To this end, we employ the regularization loss proposed by Pati *et al.* [27] that aligns specific dimensions of $z^{(\text{pln})}$ with the target expressive attributes. This method assumes that a latent representation can be disentangled through its monotonic relationship with a target attribute. Let d_i and d_j be a target dimension d of ith and jth latent representations, respectively, where $d \in z_a^{(\text{pln})}$, $i, j \in [1, M]$, and M is the size of a mini-batch. A distance matrix $\mathcal{D}_d$ is computed between d_i and d_j within a mini-batch, where $\mathcal{D}_d = d_i - d_j$. A similar distance matrix $\mathcal{D}_a$ is computed for the two target attribute values a_i and a_j. We minimize a MSE between $\mathcal{D}_d$ and $\mathcal{D}_a$ as follows:

$$\mathcal{L}_{\text{reg}} = \text{MSE}(\tanh(\mathcal{D}_d), \text{sign}(\mathcal{D}_a)) \quad (6)$$

2.5 Overall Objective

The overall objective of our proposed network aims to generate realistic performance features with properly disentangled representations for the intended factors:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda_{\text{pln}}\mathcal{L}_{\text{pln}} + \lambda_{\text{str}}\mathcal{L}_{\text{str}} + \lambda_{\text{fac}}\mathcal{L}_{\text{fac}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \quad (7)$$

where λ_{pln}, λ_{str}, λ_{fac}, and λ_{reg} are hyperparameters for balancing the importance of the loss terms.

3. EXPERIMENTAL SETUPS

3.1 Dataset and Implementation

We use Yamaha e-Competition Dataset [8] and Vienna 4x22 Piano Corpus [40]. From these datasets, we collect 356 performances of 34 pieces by Frédéric Chopin, which have been representative research subjects for analyzing the Western musical expression [6, 22, 41, 42]. We use 30 pieces (108,738 batches) for training and the rest for testing. To verify the generality of model performances, we also collect the external dataset from ASAP dataset [43]. We use 116 performances for 23 pieces by 10 composers who represent various eras of Western music. For subjective evaluation, we collect 42 songs of non-Classical songs from online source [2] which are less constrained to written expression than most Classical excerpts.

[2] http://www.ambrosepianotabs.com/page/library

We basically follow Jeong *et al.* [8] to compute the input features from the aligned pairs of performance and score data. We set MIDI velocities and Beat Per Minute (BPM) of all notes in the score data to be 64 and 120, respectively. We also remove any grace notes for simplicity and manually correct any errors. The performance features are further normalized into a range from -1 to 1 for training. We use an ADAM optimizer [44] with an initial learning rate of 1e-5, which is reduced by 5% every epoch during backpropagation. We empirically set λ_{pln}, λ_{str}, λ_{fac}, and λ_{reg} to be 1000, 100, 1, 10, respectively. We set a degree of the polynomial function computing $I^{(\text{pln})}$ as 4 through an ablation study described in the supplementary material.

3.2 Comparative Methods

To the best of our knowledge, there is no existing method that does not intentionally follow the written guidelines in the musical score. Therefore, we use variants of our proposed network as comparing methods that differ in model architecture: **Notewise** denotes the proposed model without the hierarchical learning. **CVAE** denotes a variant of Notewise where $z^{(\text{pln})}$ is substituted with the supervisory signal $I^{(\text{pln})}$. We also conduct an ablation study that investigates necessity of the four loss terms.

4. EVALUATION

We evaluate the proposed network in terms of both objective and subjective criteria.

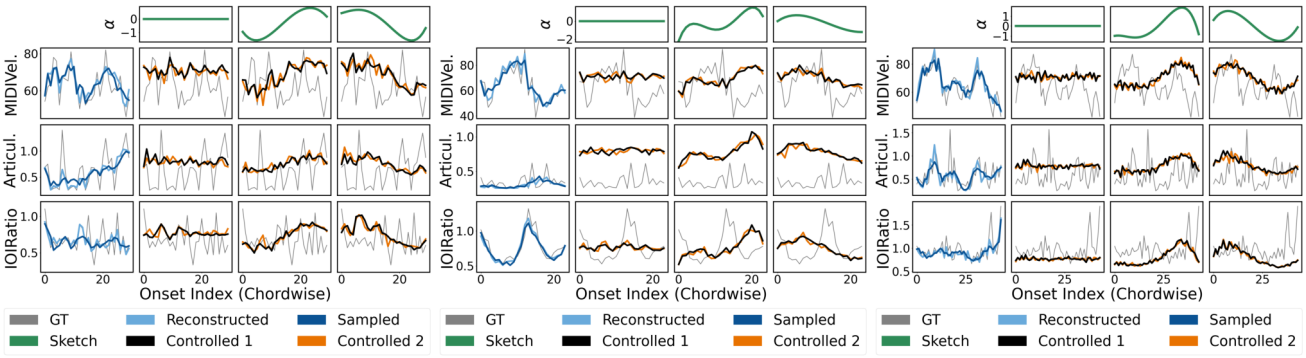
4.1 Generation Quality

We compute Pearson's correlation coefficients between the reconstructed or generated samples and human piano performances [6, 9, 11, 19]. We first measure the reconstruction quality of the test samples ("R_{recon}"). Then, we evaluate the samples generated from $\tilde{z}^{(\text{str})} \sim p_\theta(z^{(\text{str})})$ and either of : 1) $z^{(\text{pln})} \sim q_\phi(z^{(\text{pln})}|x)$ ("$\text{R}_{x|\text{pln}}$") and 2) $z_0^{(\text{pln})} \sim q_\phi(z_0^{(\text{pln})}|x_0)$ ("$\text{R}_{x|\text{pln}_0}$"), where x_0 is a zero matrix.

The results are shown in Table 1. Notewise shows the best scores in both datasets, and our method outperforms CVAE in R_{recon}. It indicates that our proposed architecture where a latent representation is used instead of a direct condition is generally good at reconstructing the human data. When using the randomly sampled $\tilde{z}^{(\text{str})}$, our method and the model without $\mathcal{L}_{\text{reg}}$ show stable scores compared to other baseline models. The model without $\mathcal{L}_{\text{reg}}$ also shows the highest scores in $\text{R}_{x|\text{pln}}$ for both datasets. It indicates that $\mathcal{L}_{\text{reg}}$ may contribute the least to generation power among other loss terms. CVAE and the model only with $\mathcal{L}^{(\text{pln})}$ also show high scores in $\text{R}_{x|\text{pln}_0}$. This may be due to the posterior collapse that makes the decoder depends mostly on the score condition [45], which is demonstrated in the supplementary material.

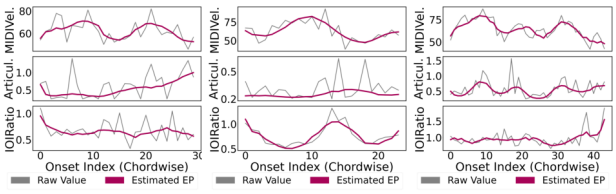
4.2 Disentangling Latent Representations

We verify whether the latent representations are well-disentangled by appropriate information [24]. To this



(a) (b) (c)

Figure 2: Qualitative samples for the proposed system. Light-blue, blue and gray lines denote the reconstructed results, sampled results from the inferred $z^{(\text{pln})}$, and their ground truths, respectively; black and orange lines denote the controlled results that are generated from different random $\tilde{z}^{(\text{str})}$; and green lines denote the "sketch" values, or α, that are inserted to $z^{(\text{pln})}$. The samples demonstrate three excerpts that are: (a) Haydn's Keyboard Sonata, Hob. XVI:39, 3rd movement, mm. 53-56; (b) Schubert's Impromptu, Op. 90, No. 4, mm. 149-152; and (c) Balakirev's Islamey, Op. 18, mm. 29-32.



(a) (b) (c)

Figure 3: Qualitative results for estimating the explicit planning from raw piano performances. Pink and gray lines denote the estimated contours and raw performance parameters, respectively. The results in (a), (b), and (c) are from the same excerpts for (a), (b), and (c) in Figure 2, respectively.

Dataset	Internal			External		
Metric	R_{recon}	$\text{R}_{x\|\text{pln}}$	$\text{R}_{x\|\text{pln}_0}$	R_{recon}	$\text{R}_{x\|\text{pln}}$	$\text{R}_{x\|\text{pln}_0}$
Notewise	**0.870**	0.392	0.203	**0.875**	0.479	0.177
CVAE	0.730	0.338	0.223	0.741	0.399	0.216
$\mathcal{L}_{\text{pln}}$	0.627	0.357	0.229	0.687	0.414	**0.220**
$\mathcal{L}_{\text{pln}} + \mathcal{L}_{\text{str}}$	0.770	0.325	0.181	0.837	0.398	0.195
w/o $\mathcal{L}_{\text{fac}}$	0.774	0.289	0.176	0.838	0.354	0.173
w/o $\mathcal{L}_{\text{reg}}$	0.737	**0.437**	0.224	0.793	**0.502**	0.216
Ours	0.737	0.427	**0.231**	0.789	0.498	0.203

Table 1: Evaluation results for the generation quality. The higher score is the better.

end, each model infers the latent representations $z^{(\text{pln})}$ and $z^{(\text{str})}$ from the test sets. Each model also randomly samples $\tilde{z}^{(\text{str})}$ and infers $z_0^{(\text{pln})} \sim q_\phi(z^{(\text{pln})}|x_0)$. We use $z_0^{(\text{pln})}$ to measure the structural attribute, since $z_0^{(\text{pln})}$ represents a flat expression where the structural attribute can be solely exposed. Each model generates new outputs as $x^{(\text{pln})} \sim p_\theta(x^{(\text{pln})}|z^{(\text{pln})}, \tilde{z}^{(\text{str})}, y)$ and $x^{(\text{str})} \sim p_\theta(x^{(\text{str})}|z_0^{(\text{pln})}, z^{(\text{str})}, y)$. Then, we compute a new signal $\tilde{I}^{(\text{pln})}$ from $x^{(\text{pln})}$ using the polynomial regression. The MSE values are calculated as $\text{MSE}_\text{p} = \text{MSE}(\tilde{I}^{(\text{pln})}, I^{(\text{pln})})$ and $\text{MSE}_\text{s} = \text{MSE}(x^{(\text{str})}, k - I^{(\text{pln})})$.

Dataset	Internal		External	
Metric	MSE_p	MSE_s	MSE_p	MSE_s
Notewise	0.003	0.006	0.022	0.028
CVAE	0.034	0.045	0.085	0.092
$\mathcal{L}_{\text{pln}}$	0.028	0.036	0.074	0.077
$\mathcal{L}_{\text{pln}} + \mathcal{L}_{\text{str}}$	0.012	0.015	0.022	0.027
w/o $\mathcal{L}_{\text{fac}}$	0.018	0.023	0.021	0.025
w/o $\mathcal{L}_{\text{reg}}$	0.002	0.004	0.014	0.022
Ours	**0.001**	**0.002**	**0.012**	**0.020**

Table 2: Evaluation results for the disentanglement of the latent representations.

Dataset	Internal			External		
Metric	C	R	L	C	R	L
Notewise	0.782	0.916	0.632	0.775	0.914	0.656
CVAE	0.798	0.812	0.620	0.773	0.802	0.649
$\mathcal{L}_{\text{pln}}$	0.693	0.852	0.323	0.694	0.834	0.324
$\mathcal{L}_{\text{pln}} + \mathcal{L}_{\text{str}}$	0.633	0.882	0.253	0.639	0.865	0.277
w/o $\mathcal{L}_{\text{fac}}$	0.831	0.846	0.789	0.832	0.831	0.847
w/o $\mathcal{L}_{\text{reg}}$	0.804	**0.955**	0.653	0.808	**0.946**	0.657
Ours	**0.942**	0.953	**0.976**	**0.944**	0.945	**0.977**

Table 3: Evaluation results for the controllability of the expressive attributes. C, R, and L denotes consistency, restrictiveness, and linearity, respectively. Each score is the average score for the expressive attributes.

Table 2 shows that our method achieves the best scores in all metrics for both datasets. This confirms that our proposed system can learn the latent representations that reflect the intended attributes. Notewise and the model without $\mathcal{L}_{\text{reg}}$ also show the robust scores compared to other baseline models. It indicates that using the notewise modeling alone is still relevant for achieving appropriate representations. It also implies that $\mathcal{L}_{\text{reg}}$ may not contribute to the disentanglement as much as other loss terms.

4.3 Controllability of Expressive Attributes

We sample a new input $\bar{x}$ where entries of each feature are constant across time. Then, each model infers

Metric	Winning Rate (Human-likeness)		
Group	T	UT	Overall
Notewise	0.317(±0.223)	0.541(±0.316)	0.493(±0.309)
CVAE	**0.467(±0.356)**	0.477(±0.342)	0.475(±0.338)
Ours	0.417(±0.256)	**0.555(±0.256)**	**0.525(±0.258)**

Table 4: Evaluation results for the winning rate in terms of human-likeness. T, UT, and Overall denote musically trained, untrained, and all groups, respectively.

$\bar{z}^{(\text{pln})} \sim q_\phi(\bar{z}^{(\text{pln})}|\bar{x})$. We control each attribute by varying dimension values of $\bar{z}^{(\text{pln})}$ following Tan *et al.* [28] and examine the new samples generated from $\bar{z}^{(\text{pln})}$. We leverage the existing metrics to measure the controllability of each model [28]: *Consistency* ("C") measures consistency across samples in terms of their controlled attributes; *restrictiveness* ("R") measures how much the uncontrolled attributes maintain their flatness over time; and *linearity* ("L") measures how much the controlled attributes are correlated with the corresponding latent dimensions. We average over the three expressive attributes–dynamics, articulation, and tempo–into one score for each metric.

Table 3 demonstrates that our system shows the best scores in consistency and linearity in both internal and external datasets. This indicates that our proposed method can robustly control the latent representation $z^{(\text{pln})}$ in intended way. The model without $\mathcal{L}_{\text{reg}}$ outperforms our method in restrictiveness. It indicates that the uncontrolled attributes by this model are the least interfered by the controlled attribute. However, its scores on consistency and linearity are lower than ours. It confirms that $\mathcal{L}_{\text{reg}}$ promotes linear control of the target attributes.

4.4 Subjective Evaluation

We conduct a listening test to compare the proposed model architecture to Notewise and CVAE. We qualitatively evaluate the base quality of the samples that have flat expressions, so that quality judgments are independent of any preference of arbitrary explicit planning. We generate each sample using $z_0^{(\text{pln})}$. A listening test is composed of 30 trials where each participant chooses a more "human-like" sample out of the generated sample and its plain MIDI [9]. Both samples have the same length which is a maximum of 15 seconds, rendered with TiMidity++ [3] without any pedal effect. *Human-likeness* denotes how similar the sample is to an actual piano performance that commonly appears in popular music. A total of 28 participants are involved, and 6 participants are professionally trained in music.

The results are demonstrated in Table 4. We measure a *winning rate*, a rate of winning over the plain MIDI, and a *top-ranking rate*, a rate of being the highest rank among the three models in terms of winning rate. These metrics are further explained in the supplementary material. The results show that musically *trained* ("T") and *untrained* ("UT") groups show the different tendency of each other: in the trained group, CVAE shows the best winning rate, and our method gets the best top-ranking rate; in the untrained group, our method shows the highest winning rate, whereas Notewise is top-ranked most frequently. We note that our system reveals smaller variances than those of CVAE and Notewise of the musically trained and untrained groups in the winning rate, respectively. Moreover, our system receives the highest overall scores for both metrics. It indicates that our system can be stably perceived more human-like than the plain MIDI compared to other baseline models.

[3] https://sourceforge.net/projects/timidity/

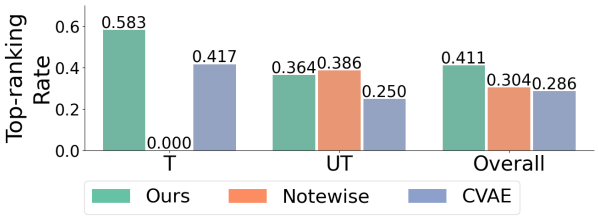


Figure 4: Evaluation results for the top-ranking rate. T, UT, and Overall denote musically trained, untrained, and all groups, respectively.

4.5 Qualitative Examples

Our system can render new piano performances from the scratch given a musical score. It can directly generate expressive parameters from the randomly sampled $\tilde{z}^{(\text{pln})} \sim p_\theta(z^{(\text{pln})})$ and $\tilde{z}^{(\text{str})} \sim p_\theta(z^{(\text{str})})$. We note that $\tilde{z}^{(\text{pln})}$ does not have temporal dependency: each $\tilde{z}_c^{(\text{pln})}$ is sampled independently of $\tilde{z}_{c-1}^{(\text{pln})}$. Hence, we need to insert specific values $\{\alpha^{(c)}\}_{c=1}^{C}$, which we call as "smooth sketches", into the target dimensions of $z^{(\text{pln})}$ if any temporal dependency of explicit planning is necessary. Figure 2 shows that the controlled parameters are greatly correlated with α, while their local characteristics follow those of the ground truth. In addition, the black and orange lines together demonstrate granular variety in the parameters induced by different $\tilde{z}^{(\text{str})}$ for the same musical structure. Moreover, Figure 3 shows that our system can estimate explicit planning from arbitrary human performances, indicating that our system can derive relevant information on explicit planning from the unseen data.

5. CONCLUSION

We propose a system that can render expressive piano performance with flexible control of musical expression. We attempt to achieve representations for the explicit planning and structural attribute through self-supervised learning objectives. We also leverage the two-step modeling of two hierarchical units for an intuitive generation. Experimental results confirm that our system shows stable generation quality, disentangles the target representations, and controls all expressive attributes independently of each other. Future work can be improving our system using a larger dataset for various genres and composers. We can also further compare our system with recent piano-rendering models [8] to investigate any connections between a performer's explicit planning and a composer's intent.

6. ACKNOWLEDGMENTS

We deeply appreciate Dasaem Jeong, Taegyun Kwon, and Juhan Nam for giving technical support to initiate this research. We also especially appreciate Hyeong-Seok Choi for providing critical feedback on the model architecture and evaluation. We greatly appreciate You Jin Choi and all of my colleagues who gave great help with respect to the listening test.

7. REFERENCES

[1] G. Widmer and W. Goebl, "Computational models of expressive music performance: The state of the art," *Journal of New Music Research*, vol. 33, no. 3, pp. 203–216, 2004.

[2] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer, "Computational models of expressive music performance: A comprehensive and critical review," *Frontiers in Digital Humanities*, vol. 5, no. 25, pp. 1–23, 2018.

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

[4] T. H. Kim, S. Fukayama, T. Nishimoto, and S. Sagayama, "Statistical approach to automatic expressive rendition of polyphonic piano music," in *Guide to Computing for Expressive Music Performance*. Springer, 2013, pp. 145–179.

[5] C. E. Cancino-Chacón and M. Grachten, "An evaluation of score descriptors combined with non-linear models of expressive dynamics in music," in *Proceedings of the International Conference on Discovery Science*, 2015.

[6] C. E. Cancino-Chacón, T. Gadermaier, G. Widmer, and M. Grachten, "An evaluation of linear and non-linear models of expressive dynamics in classical piano and symphonic music," *Machine Learning*, vol. 106, no. 6, pp. 887–909, 2017.

[7] A. Maezawa, "Deep piano performance rendering with conditional VAE," in *Late-Breaking Demo, the 19th International Society for Music Information Retrieval Conference*, 2018.

[8] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam, "VirtuosoNet: A hierarchical RNN-based system for modeling expressive piano performance," in *Proceedings of the 20th International Society for Music Information Retrieval*, 2019.

[9] D. Jeong, T. Kwon, Y. Kim, and J. Nam, "Graph neural network for music score data and modeling expressive piano performance," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013.

[11] A. Maezawa, K. Yamamoto, and T. Fujishima, "Rendering music performance with interpretation variations using conditional variational RNN," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, 2019.

[12] H. H. Tan, Y.-J. Luo, and D. Herremans, "Generative modeling for controllable audio synthesis of expressive piano performance," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[13] R. Bresin and A. Friberg, "Emotional coloring of computer-controlled music performances," *Computer Music Journal*, vol. 24, no. 4, 2000.

[14] S. R. Livingstone, R. Muhlberger, A. R. Brown, and W. F. Thompson, "Changing musical emotion: A computational rule system for modifying score and performance," *Computer Music Journal*, vol. 34, no. 1, 2010.

[15] M. Bernays and C. Traube, "Investigating pianists' individuality in the performance of five timbral nuances through patterns of articulation, touch, dynamics, and pedaling," *Frontiers in Psychology*, vol. 5, no. 157, pp. 1–19, 2014.

[16] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Computing and Applications*, vol. 32, pp. 955–967, 2020.

[17] A. Bhatara, A. K. Tirovolas, L. M. Duan, B. Levy, and D. J. Levitin, "Perception of emotional expression in musical performance," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 3, pp. 921–934, 2011.

[18] A. Friberg, R. Bresin, and J. Sundberg, "Overview of the KTH rule system for musical performance," *Advances in Cognitive Psychology*, vol. 2, no. 2-3, pp. 145–161, 2006.

[19] S. Flossmann, M. Grachten, and G. Widmer, "Expressive performance rendering with probabilistic models," in *Guide to Computing for Expressive Music Performance*. Springer, 2013, pp. 75–98.

[20] R. H. Woody, "The relationship between explicit planning and expressive performance of dynamic variations in an aural modeling task," *Journal of Research in Music Education*, vol. 47, no. 4, pp. 331–342, 1999.

[21] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "Music performance analysis: A survey," in *Proceedings of the 20st International Society for Music Information Retrieval Conference*, 2019.

[22] B. H. Repp, "A microcosm of musical expression: II. quantitative analysis of pianists' dynamics in the initial measures of Chopin's Etude in E major," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1972–1988, 1999.

[23] H. Honing, "From time to time: The representation of timing and tempo," *Computer Music Journal*, vol. 25, no. 3, 2001.

[24] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation," in *Proceedings of Computer Vision and Pattern Recognition*, 2020.

[25] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[26] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.

[27] A. Pati and A. Lerch, "Attribute-based regularization of latent spaces for variational auto-encoders," in *Neural Computing and Applications*, 2020.

[28] H. H. Tan and D. Herremans, "Music FaderNets: Controllable music generation based on high-level features via low-level feature modeling," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 2020.

[29] D. Jeong, T. Kwon, Y. Kim, and J. Nam, "Score and performance features for rendering expressive music performances," in *Proceedings of the Music Encoding Conference*, 2019.

[30] A. Roberts, J. Engel, and D. Eck, "Hierarchical variational autoencoders for music," in *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017.

[31] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[32] S.-L. Wu and Y.-H. Yang, "MuseMorphose: Full-song and fine-grained music style transfer with just one Transformer VAE," *arXiv preprint arXiv:2105.04090*, 2021.

[33] Z. Wang, Y. Zhang, Y. Zhang, J. Jiang, R. Yang, J. Z. (Jake), and G. Xia, "Pianotree VAE: Structured representation learning for polyphonic music," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[35] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

[36] Y. Li and S. Mandt, "Disentangled sequential autoencoder," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[37] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

[38] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, 1989.

[39] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[40] W. Goebl, "Melody lead in piano performance: Expressive device or artifact?" *The Journal of the Acoustical Society of America*, vol. 110, no. 1, 2001.

[41] M. Grachten and G. Widmer, "Linear basis models for prediction and analysis of musical expression," *Journal of New Music Research*, vol. 41, no. 4, pp. 311–322, 2012.

[42] Z. Shi, "Computational analysis and modeling of expressive timing in Chopin Mazurkas," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021.

[43] 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*, 2020.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.