Sketching the Expression: Flexible Rendering of Expressive Piano Performance with Self-Supervised Learning

Seungyeon Rhyu[1], Sarah Kim[2], and Kyogu Lee[1]

[1]*Music and Audio Research Group, Seoul National University, South Korea*
[2]*Krust Universe, South Korea*

ISMIR 2022 BENGALURU

Introduction

Piano performance generation: Generating parameters that fit a musical piece

- It aims to generate coherent parameters for loudness or timing, based on given musical scores.

Expanding musical creativity: Musical expression beyond the written guidelines

- Previous models for controlling a piano performance followed written expression guidelines or dealt with only partial attributes of musical expression.
- Performers can actively choose techniques to highlight various emotions or nuances, creating musical expressions beyond the guidelines.

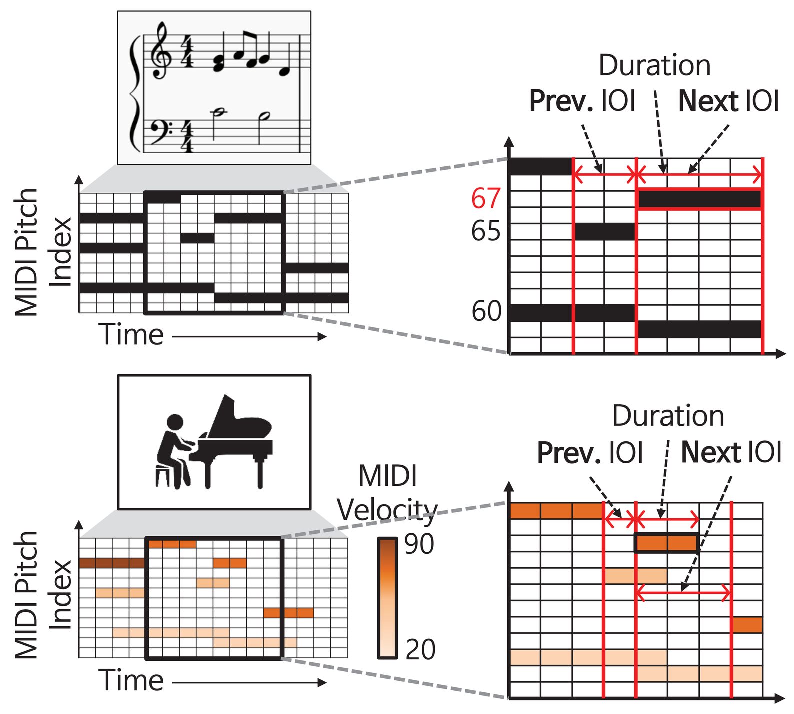
Objective: Disentangling musical expression using self-supervised learning

- We propose a generative model that disentangles two representations for high-level musical expression, or *explicit planning*, and low-level structural attributes.
- We use a conditional VAE modified for sequential data and a self-supervised learning framework to regularize the representations.

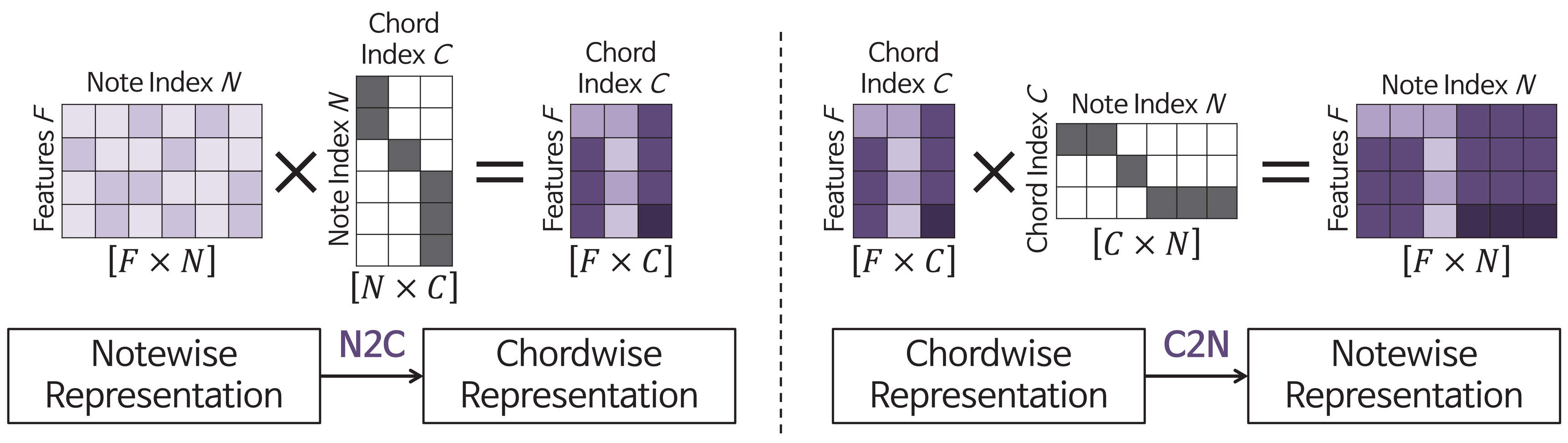
Proposed Methods

I. Data Representation

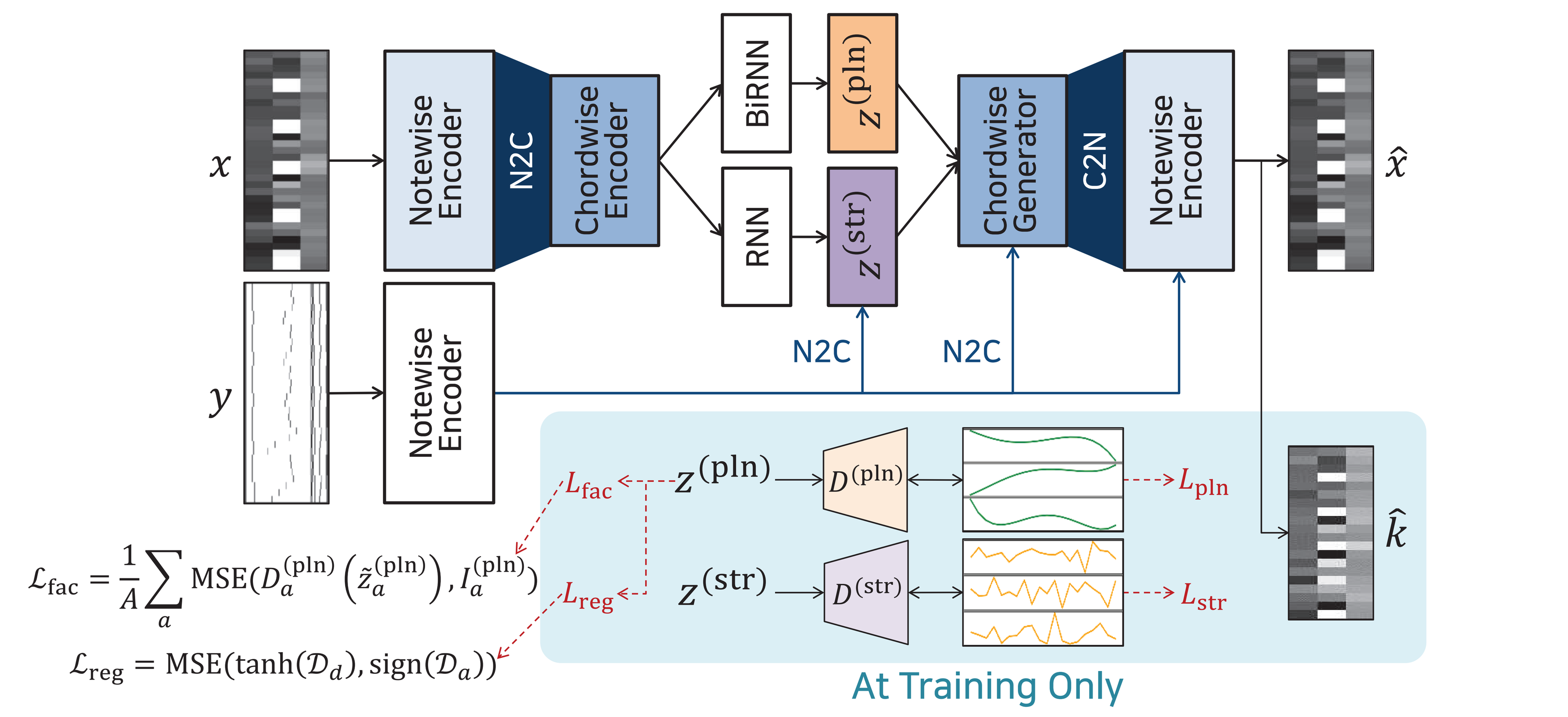
Score Feature	Performance Feature
✓ MIDI Pitch ✓ Duration (16th note) ✓ IOI (16th note) ✓ Is-top-voice ✓ #Note-in-Chord ✓ Position-in-chord ✓ Staff ✓ Is-downbeat	✓ MIDI Velocity (dynamics) ✓ IOI Ratio (tempo) $\frac{\text{Perform. Prev. IOI}}{\text{Score Prev. IOI}}$ ✓ Articulation $\frac{\left(\frac{\text{Perform. Duration}}{\text{score Duration}}\right)}{\text{IOI Ratio with Next IOI}}$



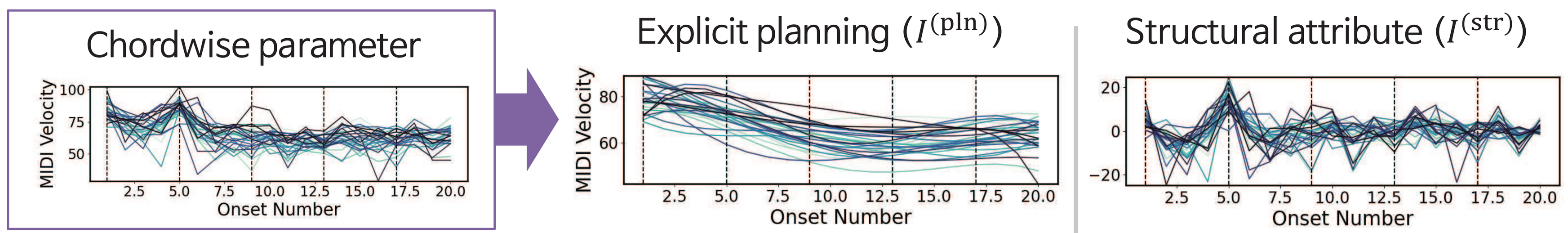
II. Modeling Musical Hierarchy



III. Model Architecture



Training Objective: $\mathcal{L} = \mathcal{L}_{\mathrm{VAE}} + \lambda_{\mathrm{pln}}\mathcal{L}_{\mathrm{pln}} + \lambda_{\mathrm{str}}\mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{fac}}\mathcal{L}_{\mathrm{fac}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}$



Computing pseudo labels from input parameter

Polynomial regression ($d = 4$) from chordwise parameter

Difference b/t chordwise parameter and $I^{(\mathrm{pln})}$

Dataset

Evaluation	Objective		Subjective
Type	Internal	External	External
Dataset	Yamaha e-Competition Vienna 4x22 Piano Corpus	ASAP	Online
Composer / Genre	Chopin only / Classical	10 composers / Classical	Various / Non-Classical
# of song/perform.	34 / 356	23 / 116	42 / (score only)

Evaluation

I. Generation Quality

Dataset	Internal			External		
Metric	$\mathrm{R}_{\mathrm{recon}}$	$\mathrm{R}_{x\vert\mathrm{pln}}$	$\mathrm{R}_{x\vert\mathrm{pln}_0}$	$\mathrm{R}_{\mathrm{recon}}$	$\mathrm{R}_{x\vert\mathrm{pln}}$	$\mathrm{R}_{x\vert\mathrm{pln}_0}$
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
L_{pln}	0.627	0.357	0.229	0.687	0.414	**0.220**
$L_{\mathrm{pln}} + L_{\mathrm{str}}$	0.770	0.325	0.181	0.837	0.398	0.195
w/o L_{fac}	0.774	0.289	0.176	0.838	0.354	0.173
w/o L_{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

- Pearson's correlation coefficients
- $\mathrm{R}_{\mathrm{recon}}$: Reconstruction loss.
- $\mathrm{R}_{x|\mathrm{pln}}$: Evaluating samples with random z_S and inferred $z^{(\mathrm{pln})}$ from data.
- $\mathrm{R}_{x|\mathrm{pln}_0}$: Evaluating samples with random z_S and inferred $z_0^{(\mathrm{pln})}$ from a zero matrix.

- Proposed architecture outperforms CVAE in most metrics.
- Proposed chordwise model generates better results with random $z^{(\mathrm{str})}$ than Notewise.
- Ours shows stable generation scores with random $z^{(\mathrm{str})}$ compared to other models.

II. Disentanglement & Controllability of Musical Expression

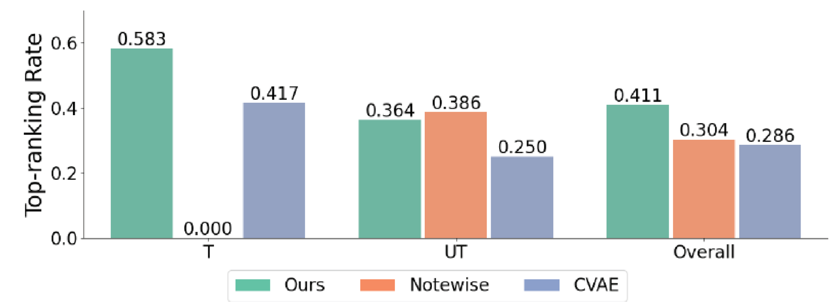
Dataset	Internal		External	
Metric	$\mathrm{MSE_P}$	$\mathrm{MSE_S}$	$\mathrm{MSE_P}$	$\mathrm{MSE_S}$
Notewise	0.003	0.006	0.022	0.028
CVAE	0.034	0.045	0.085	0.092
L_{pln}	0.028	0.036	0.074	0.077
$L_{\mathrm{pln}} + L_{\mathrm{str}}$	0.012	0.015	0.022	0.027
w/o L_{fac}	0.018	0.023	0.021	0.025
w/o L_{reg}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

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
L_{pln}	0.693	0.852	0.323	0.694	0.834	0.324
$L_{\mathrm{pln}} + L_{\mathrm{str}}$	0.633	0.882	0.253	0.639	0.865	0.277
w/o L_{fac}	0.831	0.846	0.789	0.832	0.831	0.847
w/o L_{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**

- **C:** Consistency of the controlled attribute. / **R:** Restrictiveness of the uncontrolled attribute.
- **L:** Linearity b/t the controlled attribute and corresponding latent dimension.
- Our model shows the best scores in most metrics for disentanglement & controllability.

III. Listening Test

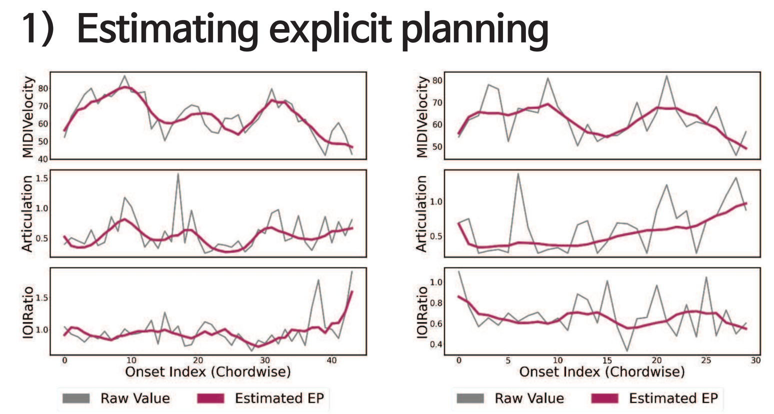
Metric	Winning Rate (Human-likeness)		
Group	T	UT	Overall
Notewise	0.317(±0.223)	0.541(±0.316)	0.493(±0.309)
CVAE	**0.467(±0.356)**	0.477(±0.342)	0.475(±0.338)
Ours	0.417(±0.256)	**0.555(±0.256)**	**0.525(±0.258)**



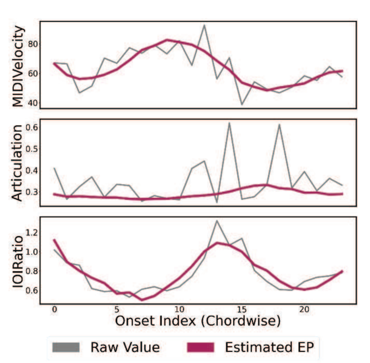
- Winning rate: a rate of winning plain MIDI (A/B test)
- Top-ranking rate: a rate of being the highest rank in winning rate.
- T/UT: musically trained (6) / untrained (22)

III. Quantitative Results

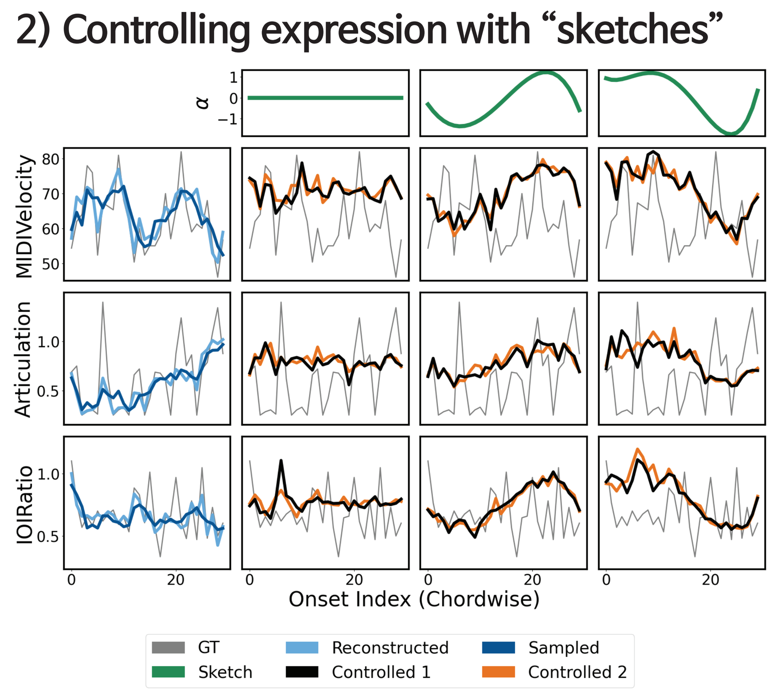
1) Estimating explicit planning



- Estimated explicit planning from $z^{(\mathrm{pln})}$ well fits the original values of the expressive attributes.



2) Controlling expression with "sketches"



- α: a sequence of values ("sketches") fed to a latent dimension $z^{(\mathrm{pln})}$ for controlling a target expressive attribute.

Baseline Methods

I. Architecture

	Description
Notewise	Ours without chordwise encoding and decoding
CVAE	Variant of Notewise where $z^{(\mathrm{pln})}$ is substituted with the supervisory signal $I^{(\mathrm{pln})}$

II. Ablation Study

	Description
w/ $\mathcal{L}_{\mathrm{pln}}$	Ours only with prediction task using $z^{(\mathrm{pln})}$
w/ $\mathcal{L}_{\mathrm{pln}} + \mathcal{L}_{\mathrm{str}}$	Ours with prediction tasks using $z^{(\mathrm{pln})}$ and $z^{(\mathrm{str})}$
w/o $\mathcal{L}_{\mathrm{fac}}$	Ours without the additional factorization loss
w/o $\mathcal{L}_{\mathrm{reg}}$	Ours without regularization method* for sketch-control

*A. Pati and A. Lerch. 2019. Latent space regularization for explicit control of musical attributes. *In Proceedings of the 36th International Conference on Machine Learning.*

Conclusion

Piano performance rendering with flexible musical expression

- Our proposed system disentangles entire musical expression from piano performance and flexibly renders expressive piano performances in stable quality.
- Dynamics, articulation, and tempo can be independently controlled by our system while other structural attributes maintain their state.

Future work

- Deeper investigation for computing $I^{(\mathrm{pln})}$ with other possible methods.
- Outputs can be rendered from scratch with random $z^{(\mathrm{pln})}$ and $z^{(\mathrm{str})}$. However, the random $z^{(\mathrm{pln})}$ does not inherit temporal dependency without given sketch. Future study is needed for inferring $z^{(\mathrm{pln})}$ that has temporal dependency without any specific sketch given as the input.