DISENTANGLING THE HOROWITZ FACTOR: LEARNING CONTENT AND STYLE FROM EXPRESSIVE PIANO PERFORMANCE

Huan Zhang, Simon Dixon

Centre for Digital Music, Queen Mary University of London, UK

ABSTRACT

In the Western art music tradition, expressive piano performance consists of two kinds of information: the score, with pitch and timing expressed in simple musical units along with occasional expression instructions, and the performer's interpretation of the score, involving variations in tempo, dynamics and articulation. In this paper, we present a novel framework for learning representations that disentangle musical content and performance style from expressive piano performances in an unsupervised manner. Our method is based on an extension of the vector-quantized variational autoencoder (VQ-VAE) with individual content and style branches, along with mutual information (MI) minimization techniques and self-supervising strategies. We performed experiments and ablation studies on the ATEPP dataset, a large set of automatically transcribed virtuosic piano performances with rich stylistic variations, and evaluated the content reconstruction and style discrimination in a style-transfer manner. Our experiments demonstrate that the model learnt separate latent variables that encode musical content (such as pitch and relative timing) and stylistic attributes, as generated samples align well with the content input with low note error rates (NER), and the 40-way style discrimination proxy task outperformed the baseline with top1 accuracy of 0.168.

***Index Terms*—** Content-style disentanglement, representation learning, style transfer, expressive piano performance

1. INTRODUCTION AND RELATED WORK

Expressive music performance is the art of shaping a musical piece by continuously varying interpretative parameters such as tempo and dynamics. Human musicians do not play a piece of music mechanically as written in the printed music score. When we hear a performance, two pieces of information are heard: the conceptualized composition described by strict musical units with occasional, general expression instructions, and the performer's interpretative input that consists of variations like speeding up, slowing down, stressing certain notes or passages, and so on. More importantly, such artistic decisions are often highly specific to each individual performer, and there have been numerous attempts [1, 2] to characterise the individual styles of performers (e.g., the so-called "Horowitz factor" [3]). In Fig. 1, a visualization [1] is shown to demonstrate such expressive factors.

Such division between structure and aesthetics has also manifested in other domains: in visual arts and general image processing, geometric information can be distilled and isolated from textural properties [4]; in speech and audio, voice and speaker information is learnt separately from the text content. Understanding and disentangling such components leads to applications like image completion [5], artistic style transfer [6], and speech synthesis [7]. However, few researchers attempt to disentangle content and style for expressive music performance.

Fig. 1. A visual comparison between the metrical score (middle) and Brendel's expressive rendering (bottom) of the opening phrase of Mozart's K310 sonata slow movement (top). Expressive devices such as the asynchrony of the chords and *ritardando* towards the end of phrase are clearly identifiable.

Previous work on music transformation using audio data [8, 9] focuses on isolating timbre from pitch in a similar fashion as in speech transfer learning. For work in the symbolic music realm [10, 11, 12], the focus usually lies in disentangling aspects of the compositional content such as harmony, texture and arrangement, especially with the aim of controllable music generation. Moreover, unlike compositional disentanglement works that utilize pop music MIDI and MusicXML datasets with annotations of chords and meter [10, 13], our work requires expressive performance data that contains simpler (no annotations) but more nuanced information (such as precise onset timing and pedals) [14], and thus brings higher complexity to the model.

Meanwhile, many methods have been applied to learn the disentangled representation of style and content from data. Under the variational autoencoder (VAE) framework, models like FH-VAE [15], DSAE [16], TS-DSAE [9] have been proposed for encoding and generating high dimensional sequential data. We also followed this framework of encoder-decoder structure, but the usual approach based on decoupling *global* and *local* tokens [9, 17] does not align well with our task, since unlike voice or timbre which can be summarized at the sequence level, expressive deviation is not a global but a time-varying attribute. Other work on disentanglement is based on generative adversarial networks (GANs) [18], but they can be hard to train and require careful hyperparameter tuning [19]. Meanwhile, various techniques have been applied in guiding the model for disentanglement, such as minimizing the mutual information between latent variables [20, 7] and adversarial training [6]. Another viable strategy is to introduce additional information such as the chord progression reconstruction [10].

[1] Generated by https://midialignment.github.io/score-performance-match-editor/ScorePerformanceMatchEditor.html

To our knowledge, this is the first work that address music style translation from a performance interpretation perspective. Our contributions can be summarized as follows:

- We present the first neural framework for learning content and style representation in expressive piano performance.
- We propose new evaluation metrics for this specific task, such as NER for validating the content reconstruction and a proxy performer recognition task for style discrimination.
- Using a dataset [14] of 11742 transcribed classical piano performances with rich stylistic variation, our model learnt separate latent representations in an unsupervised manner, outperforming the baseline in both style and content evaluations.

2. METHODOLOGY

2.1. Problem Formulation and Loss Objectives

Based on the assumption that each performance rendering is a combination of musical content and interpretative input, the likelihood of observing the performance sample $\boldsymbol{X}$ given content information $\boldsymbol{Z}_c$ and style information $\boldsymbol{Z}_s$ is $p_\theta(\boldsymbol{X}|\boldsymbol{Z}_c, \boldsymbol{Z}_s)$, where θ is the model parameters. In VAE, we use variational inference to learn an approximate posterior for each latent variable through encoder functions $q_c(\boldsymbol{Z}_c|\boldsymbol{X})$ and $q_s(\boldsymbol{Z}_s|\boldsymbol{X})$, with optimization proved by evidence lower bound (ELBO). The base loss function for the two-branch VAE is shown in Eq. 1, with reconstruction and Kullback-Leibler (KL) divergence.

$$\mathcal{L}_{base} = \mathbb{E}_{p(X)}\mathbb{E}_{q_\theta(X|Z_S,Z_C)}[-\log(p_\theta(X|Z_S,Z_C))] + \mathbb{E}_{p(X)}[\mathrm{KL}(p(Z_S)||q_\theta(Z_S|X))] \quad (1)$$

Mutual Information Minimization We followed the MINE [21] method to construct a lower bound of mutual information based on the Donsker-Varadhan representation of KL divergence as shown in Eq. 2. By minimizing the mutual information $I(\boldsymbol{Z}_c, \boldsymbol{Z}_s)$ between the hidden representations $\boldsymbol{Z}_c$ and $\boldsymbol{Z}_s$ which equals to the divergence of their joint distribution $P_{(\boldsymbol{Z}_c,\boldsymbol{Z}_s)}$ and product of marginals $P_{\boldsymbol{Z}_c} \times P_{\boldsymbol{Z}_s}$, we alleviate possible content leakage and ensure disentanglement. In the equation, the supremum is taken over all functions G such that the two expectations are finite. Given that there is no closed-form computation of mutual information, we use a neural network G to approximate this lower bound of mutual information, and it is optimized along with the main network.

$$\mathcal{L}_{MI} = \sup_G \mathbb{E}_{P_{(\boldsymbol{Z}_c,\boldsymbol{Z}_s)}}[G] - \log(\mathbb{E}_{P_{\boldsymbol{Z}_c}\times P_{\boldsymbol{Z}_s}}[e^G]) \quad (2)$$

Vector Quantization The technique of vector quantization (VQ) [22] has been proven effective in multiple disentanglement tasks [23]. The VQ layer encourages the content encoder output $z_e(x)$ to minimize the distance between itself and the nearest codebook vector e. The VQ loss in Eq. 3 is added, where $\mathrm{sg}(\cdot)$ is the stop gradient operation. In our experiments, we take the commitment loss weight α as 1.

$$\mathcal{L}_{VQ} = \|\mathrm{sg}[z_e(x)] - e\|_2^2 + \alpha\|z_e(x) - \mathrm{sg}[e]\|_2^2 \quad (3)$$

Our overall loss objective is comprised of the above elements, where β_1 and β_2 are weighting parameters:

$$\mathcal{L} = \mathcal{L}_{base} + \beta_1\mathcal{L}_{VQ} + \beta_2\mathcal{L}_{MI} \quad (4)$$

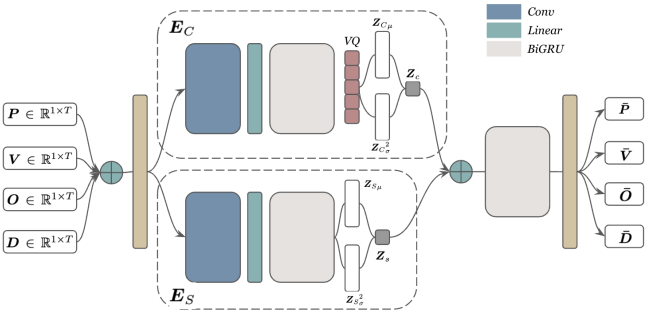

Fig. 2. Model architecture.

2.2. Input Representation

Each piece of symbolic data is represented by four token sequences $\boldsymbol{P} \in \mathbb{R}^{1\times T}, \boldsymbol{V} \in \mathbb{R}^{1\times T}, \boldsymbol{O} \in \mathbb{R}^{1\times T}, \boldsymbol{D} \in \mathbb{R}^{1\times T}$, corresponding to pitch, velocity, onset and duration. The four sequences are each fed through an embedding layer and then concatenated into input $\mathbf{X} \in \mathbb{R}^{T\times embDim}$, similar to the compound word (CP) symbolic music tokenization scheme [24]. In inference, four separate projection layers invert this process and output generated token sequences $\bar{\boldsymbol{P}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{O}}, \bar{\boldsymbol{D}}$.

Vocabulary-wise, following the MIDI standard, $\boldsymbol{P}$ and $\boldsymbol{V}$ both take on 128 values, and $\boldsymbol{O}$ and $\boldsymbol{D}$ take on 2300 values and 700 values, respectively. The time tokens are quantized with 10ms resolution, and we take sequence length $T = 128$.

Although the MIDI vocabularies of $\boldsymbol{P}, \boldsymbol{V}, \boldsymbol{O}, \boldsymbol{D}$ are discrete, they are not actually categorical as they have continuous semantic meaning of pitch and timing. Thus, in terms of reconstruction loss we experimented with regression into the output with the L2 norm loss instead of cross-entropy classification, so that the distances between vocabulary classes are incorporated into training.

2.3. Model Architecture and Training Details

Our overall model architecture is summarized in Fig. 2. As described in section 2.2, the symbolic music input and output sequences are processed via an embedding layer and a projection layer, respectively, from their tokenized representation of MIDI events.

The content encoder $\boldsymbol{E}_C(\cdot)$ aims to extract a sequence of latent variables $\boldsymbol{Z}_c \in \mathbb{R}^{T\times LatentDim}$ that only represent the content from the input $\boldsymbol{X}$. The content encoder is built on top of a convolutional stack and two layers of bidirectional gated recurrent units (GRU) to represent the musical content in a context-aware fashion. As mentioned in section 2.1, the information bottleneck is applied on top of the content encoder via a vector quantization layer with a codebook size of 4096, guiding the branch to focus on localized information.

$\boldsymbol{E}_S(\cdot)$ functions as the style branch in our architecture, and aims to factor out the style latent that only represents the expressive deviations. It is built with a similar architecture, but without the VQ layer. Both branches have a variational layer at the end and the latents are sampled according to $\boldsymbol{Z}_\mu$ and $\boldsymbol{Z}_\sigma^2$.

We train the model using Adam to minimize the loss from Eq. 4. We trained for 450 epochs, taking about 46 hours in total on two RTX 2080 GPUs. We take $embDim = 128$ and $LatentDim = 512$, and for the loss weighting parameters, we used $\beta_1 = 0.5$ and $\beta_2 = 0.5$. Ablation studies on other parameters are presented in the results section.

Baseline Given the limited prior work on our topic, we set our baseline as the vanilla VAE framework with the loss objective described in Eq. 1.

Self-Supervised Training Inspired by Cífka et al. [8], we also explore the self-supervised training technique. To ensure that the style encoder only encodes style, we feed the style encoder another segment $\boldsymbol{X}_j$ from the same training set recording as the content input $\boldsymbol{X}_i$. The rationale is that given the same expressive style throughout a recording, even if $\boldsymbol{X}_j$ has different content, the model should be able to reconstruct $\boldsymbol{X}_i$ with the style latent from $\boldsymbol{X}_j$ and content latent from $\boldsymbol{X}_i$. Besides the paired segment input, other training objectives and the model architecture remain the same as for the main experiment.

3. EXPERIMENTS

3.1. Dataset

The content and style experiments are supported by the ATEPP dataset [14], which contains 11742 tracks of virtuosic solo piano performances in MIDI format obtained via automatic transcription. The transcribed MIDI files contains detailed expressive information such as the key velocity and pedal depths. Also, as the data are in the symbolic domain, piano acoustic differences can be reasonably ignored in our context. With 49 pianists performing an overlapping corpus of standard Western classical repertoire, rich stylistic variations are represented in this dataset. The training segments and input representations are generated following the procedure in Sec. 2.2. We split the data into train/valid/test sets by each track of music instead of individual segments, as repetition in the music might otherwise compromise the test set.

In this project, we simplify the labelling of expressive style by using performer identity as a proxy. We acknowledge that from a musical perspective, there does not exist a bijective mapping between performer and interpretation style. But given the subjective nature of interpretation, very few objective parameters of performance style have been proposed [25], so this is a reasonable approximation.

3.2. Evaluation

We evaluate the effectiveness of our disentanglement model from a style-transfer perspective. In test-time generation, the decoder takes a content input $\boldsymbol{X}_c$ and a style input $\boldsymbol{X}_s$ from a different excerpt, concatenates their hidden representations and decodes an expressive rendering $\hat{\boldsymbol{X}}$. Considering effects that the proximity of inputs may have on the results, the following input shuffling schemes are proposed:

1. **SR**: X_c and X_s are taken from the **S**ame **R**ecording
2. **SD**: **S**ame performer but **D**ifferent piece
3. **DP**: Different recordings from **D**ifferent **P**erformers

At test-time, a set of samples is generated for evaluation of each scheme, by selecting inputs X_c and X_s according to the respective scheme.

Content Preservation: For evaluating the faithfulness of content reconstruction, we introduce the measure of note error rate (NER) that is analogous to the word error rate (WER) used in speech recognition [7, 20]. An alignment of the generated $\hat{\boldsymbol{X}}$ and content input $\boldsymbol{X}_c$ is produced by Nakamura's algorithm [26]. This algorithm employs hidden Markov models (HMMs) to align two symbolic performances and correct errors. The NER is then calculated from the alignment outputs, where S_{extra}, S_{wrong} and $S_{missing}$ denote the set of extra, pitch-incorrect and missing notes of the generated MIDI performance with regards to the content input returned by the alignment function.

$$NER = \frac{|S_{extra}| + |S_{wrong}| + |S_{missing}|}{|S_{matched}| + |S_{missing}|} \tag{5}$$

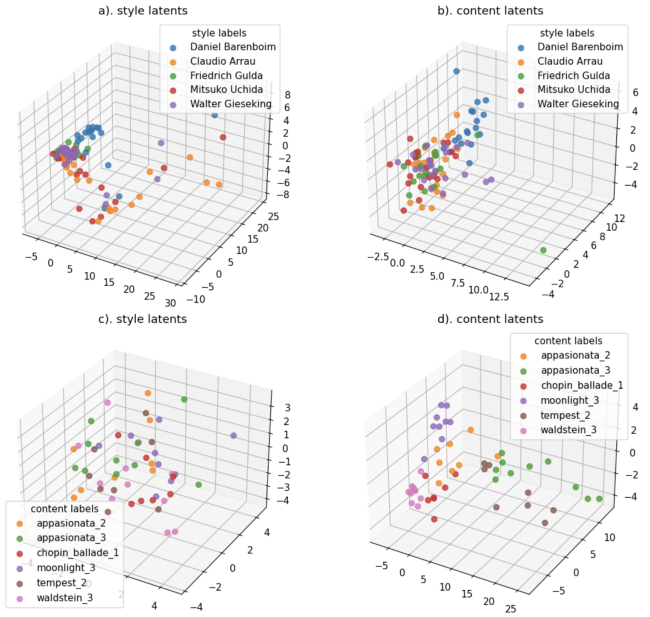

Fig. 3. Visualization of latent variables, showing greater proximity of performers by style (a) than content (b), and of pieces by content (d) than style (c).

Style Fit As mentioned in section 3.1, stylistic characterization is subjective and no standardized measure exists. Thus, we achieve the evaluation for style fitness via neural approximation. A neural network discriminator D, acting as a probe [27], is trained to evaluate how well the generated samples simulate the ground-truth distribution [28]. D is first trained on generated samples and then discriminates on ground truth test data as a 40-way classification task of style discrimination (a few pianists are not present in the test split). D is a simple recurrent neural network consisting of an embedding layer, 2 layers of biGRUs and a softmax projection. The discriminator is trained on the generated data for 300 epochs with an early stopping of 10 epochs to prevent overfitting. Top1 and Top5 accuracy are reported.

3.3. Results and Discussions

Table 1 shows the results of our experiments. Both proposed models PERFVAE and PERFVAE$_{SS}$ outperform the vanilla VAE baseline. In terms of NER, both models achieved less than 0.2, which means the generated content is roughly 80 percent aligned with the desired content. In terms of style discrimination, on the 40-way classification task we achieved the highest accuracy of 0.168, demonstrating the style-transferred generative output partially matched the ground-truth style distribution. With the content and style evaluation results combined, we can informally say that the disentanglement is partially successful and can be viewed as a starting point for this novel task.

Configuration	Valid. recon. acc.	Shuffle	*Content* Note error rate	*Style* Top1 acc.	*Style* Top5 acc.
BASELINE	0.875 ± 0.087	SR	0.234 ± 0.036	0.102 ± 0.032	0.281 ± 0.047
		SD	0.241 ± 0.036	0.067 ± 0.051	0.223 ± 0.063
		DP	0.268 ± 0.075	0.045 ± 0.033	0.212 ± 0.046
PERFVAE	0.726 ± 0.094	SR	**0.121 ± 0.046**	**0.168 ± 0.051**	0.341 ± 0.041
		SD	0.123 ± 0.038	0.145 ± 0.050	0.304 ± 0.035
		DP	0.171 ± 0.051	0.098 ± 0.037	0.281 ± 0.037
PERFVAE$_{SS}$	0.713 ± 0.102	SR	0.166 ± 0.033	0.164 ± 0.030	**0.347 ± 0.052**
		SD	0.172 ± 0.056	0.151 ± 0.042	0.307 ± 0.060
		DP	0.201 ± 0.033	0.130 ± 0.061	0.263 ± 0.047

Table 1. Comparison of different methods in terms of both content and style measures with 0.95 confidence. PERFVAE is the proposed model with the loss objective from Eq. 4; PERFVAE$_{SS}$ is the proposed model with self-supervised training as described in Sec. 2.

We also note that the self-supervised model performs less accurately in NER (content reconstruction) than the unsupervised version, which is possibly due to the fact that different content goes through the style branch during the training process. In terms of the reconstruction accuracy regarding the input original, the proposed models actually perform worse than the vanilla VAE baseline. This is possibly due to the fact that more regularization is placed on the proposed models' training objective.

Another interesting observation involves different shuffle groups. The results for both content and style measures show a decrease from SR group to DP group. Given that the groups correspond to high proximity and low proximity respectively of the pair of inputs, we can infer that the model struggles as the content and style inputs become more distant and less musically plausible (for example, blending an Ashkenazy performance of a Chopin ballad with Gould's Bach Inventions).

Our subjective observations upon examining the outputs mostly match the objective evaluation. We find that the musical content from content input is generally well-preserved in the style-transferred output. Also, under the shuffle group DP, the output is more disorganized compared to the other two groups, demonstrating that the disentanglement quality is still quite limited. We also demo a subset of generated examples.[2]

Ablation Study We also performed an ablation study on the effect of VQ codebook size as well as the use of mutual information loss. As shown in table 2, the incorporation of mutual information loss helped the model produce better disentangled results in both content and style measures in all configurations. There are some positive correlation between increasing codebook size and decreasing NER results as well as discrimination accuracy, but not too much significance was observed, even when we increase the codebook size to 8192. This might be attributed to the codebook collapse [29] issue that is common in VQ-VAE.

3.4. Latent Space Analysis

In Fig. 3, we analyze the information content learnt in latent variables $\boldsymbol{Z}_C$ and $\boldsymbol{Z}_S$ by projecting them into the first three principle components using PCA. We prepared a set of data samples, which are created from combinations taken from five different styles and five different musical excerpts. In Fig. 3a and 3b, the colors are based on style labels, and in 3c and 3d the colors are based on musical content. In the dimension-reduced latent space, the style latents from the data points from the same style label have style latents grouped more closely together (Fig. 3a) than their content latents (3b). Similarly in the bottom two plots for data points containing the same musical content, no correlation of style latents is observed (3c) but the content latents show some clustering (3d).

[2]https://tinyurl.com/csd-examples

Codebook size	MI loss	NER(SR)	Top1-acc(SR)
2048	✓	0.115 ± 0.047	0.151 ± 0.037
	×	**0.109 ± 0.064**	0.141 ± 0.049
4096	✓	0.121 ± 0.046	**0.168 ± 0.051**
	×	0.116 ± 0.033	0.134 ± 0.041
8192	✓	0.126 ± 0.051	0.162 ± 0.045
	×	0.118 ± 0.060	0.140 ± 0.039

Table 2. Results of ablation studies on the codebook size and mutual information loss, all performed on the SR shuffle group without the self-supervising strategy.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework for content and style disentanglement for expressive piano performance. Under the vector-quantized variational autoencoder architecture with mutual information minimization, our model demonstrated effective decoupling of musical content information and performance style. Unlike previous work, we demonstrate the feasibility of unsupervised learning of expressive performance data without score annotation, thus enabling much larger-scale analysis of performance style. We hope this work can shed light on the realm of expressive performance understanding, especially on the relationship between composition elements and interpretative inputs.

In the future, we plan to extract more musically-grounded features by guiding the training, as well as setting up a more standardized profile for style characterization evaluation with the support of subjective assessment.

5. ACKNOWLEDGEMENT

This work is supported by the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, funded by UK Research and Innovation [grant number EP/S022694/1].

6. REFERENCES

[1] Gerhard Widmer and Patrick Zanon, "Automatic recognition of famous artists by machine," *Frontiers in Artificial Intelligence and Applications*, vol. 110, pp. 1109–1110, 2004.

[2] Asmir Tobudic and Gerhard Widmer, "Learning to play like the great pianists," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, pp. 871–876.

[3] Gerhard Widmer, Simon Dixon, Werner Goebl, Elias Pampalk, and Asmir Tobudic, "In search of the Horowitz factor," *AI Magazine*, vol. 24, no. 3, pp. 111–130, 2003.

[4] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy, "Disentangling content and style via unsupervised geometry distillation," in *International Conference on Learning Representations Workshop Proceedings*, 2019.

[5] Andrew Gilbert, John Collomosse, Hailin Jin, and Brian Price, "Disentangling structure and aesthetics for style-aware image completion," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Björn Ommer, "Content and style disentanglement for artistic style transfer," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[7] Ting Yao Hu, Ashish Shrivastava, Oncel Tuzel, and Chandra Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[8] Ondřej Cífka, Alexy Ozerov, Umut Şimşekli, and Gael Richard, "Self-supervised VQ-VAE for one shot music style transfer," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[9] Yin-Jyun Luo, Sebastian Ewert, and Simon Dixon, "Towards robust unsupervised disentanglement of sequential data: A case study using music audio," *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[10] Ziyu Wang, Dingsu Wang, Yixiao Zhang, and Gus Xia, "Learning interpretable representation for controllable polyphonic music generation," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[11] Ashis Pati and Alexander Lerch, "Is disentanglement enough? On latent representations for controllable music generation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[12] Ondřej Cífka, Umut Şimşekli, and Gaël Richard, "Supervised symbolic music style translation using synthetic data," *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 588–595, 2019.

[13] Yun Ning Hung, I. Tung Chiang, Yi An Chen, and Yi Hsuan Yang, "Musical composition style transfer via disentangled timbre representations," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[14] Huan Zhang, Jingjing Tang, Syed Rafee, Simon Dixon, and George Fazekas, "ATEPP: A dataset of automatically transcribed expressive piano performance," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[15] Wei Ning Hsu, Yu Zhang, and James Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems*, 2017.

[16] Yingzhen Li and Stephan Mandt, "Disentangled sequential autoencoder," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[17] Jiachen Lian, Chunlei Zhang, and Dong Yu, "Robust disentangled variational speech representation learning for zero-shot voice conversion," *Proceedings of the International Conference on Acoustics, Speech and Signal Processsing (ICASSP)*, pp. 6572–6576, 2022.

[18] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser M. Nasrabadi, "Style and content disentanglement in generative adversarial networks," *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision, WACV*, 2019.

[19] Mario Lucic, Karol Kurach, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly, "Are GANs created equal? A large-scale study," *Advances in Neural Information Processing Systems*, 2018.

[20] Andros Tjandra, Ruoming Pang, Yu Zhang, and Shigeki Karita, "Unsupervised learning of disentangled speech content and style representation," *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[21] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm, "Mutual information neural estimation," *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[22] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, 2017.

[23] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," *Proceedings of the 39th International Conference on Machine Learning (PMLR)*, 2022.

[24] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[25] Efstathios Stamatatos and Gerhard Widmer, "Automatic identification of music performers with learning ensembles," *Artificial Intelligence*, vol. 165, pp. 37–56, 2005.

[26] Eita Nakamura, Kazuyoshi Yoshii, and Haruhiro Katayose, "Performance error detection and post-processing for fast and accurate symbolic music alignment," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[27] Guillaume Alain and Yoshua Bengio, "Understanding intermediate layers using linear classifier probes," in *International Conference on Learning Representations Workshop Proceedings*, 2017.

[28] Timothee Lesort, Andrei Stoian, Jean-Francois Goudou, and David Filliat, "Training discriminative models to evaluate generative ones," in *Proceedings of the International Conference on Artificial Neural Networks*, 2019.

[29] Sander Dieleman, Aäron Van Den Oord, and Karen Simonyan, "The challenge of realistic music generation: Modelling raw audio at scale," *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.