Hierarchical Symbolic Music Generation with Variational Autoencoder-based Bar-wise Feature Sequences

Keito Sawada, Wen-Chin Huang and Tomoki Toda
Nagoya University, Japan
E-mail: sawada.keito@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—This paper proposes a hierarchical symbolic music generation method based on autoregressive modeling of variational autoencoder (VAE)-based bar-wise features. The method represents music as a sequence of low-dimensional bar-level features, enabling efficient modeling of long-term structure while maintaining local coherence. It consists of a VAE-based encoder and decoder for bar-wise feature extraction and composition, and a Transformer-based feature sequence generator conditioned on chord progression. To evaluate global structural coherence, we introduce a new metric, Bar-wise Feature Similarity Distance (BFSD). Experimental results show that the proposed method improves long-term structure compared to baseline models and achieves local naturalness comparable to existing methods. The source code for BFSD is available at https://github.com/KateSawada/barwise_feature_similarity_distance.**

I. Introduction

Symbolic music generation (SMG) aims to compose music in symbolic formats such as MIDI, REMI-based tokens [1], or piano-roll representations [2]. A key requirement for musically satisfying compositions is balancing predictability and surprise, often achieved through the interplay between local and global structures [3], [4]. Music exhibits a hierarchical structure, where small units (e.g., notes, bars) are nested within larger forms (e.g., phrases, sections) [5]–[7]. Modeling this hierarchy is essential for generating coherent music across local and global time scales.

Transformer-based autoregressive models [8], such as Music Transformer [9], MuseNet [10], and Pop Music Transformer [1], have shown strong performance by capturing long-range dependencies via self-attention. However, their reliance on fine-grained symbolic representations such as note-level tokens leads to long sequences, which increases computational cost and limits efficiency. Moreover, they typically require large datasets to learn long-term structure effectively.

To overcome these limitations, this paper proposes a hierarchical SMG method based on autoregressive modeling of variational autoencoder (VAE)-based bar-wise features. By representing music as sequences of low-dimensional bar-level features, our method reduces sequence length and allows the model to focus on global structure. It is conditioned on chord progression information, allowing the model to reflect the given harmonic context in the generated music. Explicit modeling of bar-wise feature sequences may also improve structural coherence, though melodic expressiveness remains a challenge. In addition, we introduce a new metric, Bar-wise Feature Similarity Distance (BFSD), to evaluate global structural coherence by quantifying similarity patterns among bar-wise features. Experimental results show that our method improves long-term structural consistency, as measured by BFSD metric, while maintaining local naturalness comparable to existing methods. However, subjective evaluations highlight limitations in creativity and variation, motivating further refinement.

The main contributions of this work are as follows:

- We propose a hierarchical generation framework that models music as a sequence of VAE-based bar-wise features.
- We introduce BFSD, a new objective metric for quantifying global structural coherence based on bar-wise similarity patterns.
- We demonstrate improved global structure generation, as measured by our proposed metric, BFSD, while achieving local musical naturalness comparable to existing methods.

II. Related Work

A. Modeling Local Musical Structures with Fixed-Length Representations

One of the key tasks in SMG is modeling local musical structures within fixed-length segments, such as bars or short phrases. These approaches often employ generative models such as Variational Autoencoders (VAE) [11], Generative Adversarial Networks (GAN) [12] and Diffusion Models [13], [14], which learn to produce musically plausible patterns within constrained contexts [15]–[20].

For example, MusicVAE [15] utilizes a hierarchical VAE framework to generate coherent latent representations of musical phrases, supporting interpolation and variation. Polyffusion [17] employs a diffusion-based approach to generate fixed-length piano-roll segments, conditioned on chord progressions and texture embeddings, allowing control over harmony and texture.

Recent advancements in these models have achieved near-human quality in short music generation, showing stylistic and harmonic coherence. However, applying such models to full-length compositions is challenging due to increased computational cost and the need for larger datasets.

979-8-3315-7206-8/25/$31.00

B. Modeling Global Musical Structures

For full-length generation, autoregressive models, especially Transformer-based ones, have shown success in capturing long-term dependencies [1], [9], [21], [22]. They model event-level temporal structure and enable coherent generation across extended time spans.

Music Transformer [9] introduced relative positional encoding to enhance global coherence by removing absolute position information. MuseMorphose [21] combined a VAE with a Transformer to allow fine-grained style transfer via segment-level controls such as rhythmic intensity and polyphony. Museformer [22] introduced bar-level summarization and multi-scale attention mechanisms for efficient modeling of both short- and long-term dependencies, achieving coherent long-form generation. These Transformer-based methods have contributed to improving long-range coherence, rhythmic structure, and stylistic expressiveness.

Beyond autoregressive models, alternative methods have emerged. A cascaded diffusion model for full-length SMG [23] models hierarchical structure across multiple levels. It enables coherent composition without token-level generation by iteratively refining global form while preserving local consistency. Such approaches offer promising alternatives to address the limitations of autoregressive models.

Despite these advances, challenges remain. Sequential models often struggle with fine-grained local control, as they prioritize long-range structure. Moreover, long sequences demand high computational resources, limiting scalability. Fixed-length generation methods also face difficulties when extended to longer compositions, often degrading in performance due to data and resource constraints.

These challenges underscore the need for methods that model global structure efficiently while retaining local control—an issue this work aims to address.

III. Proposed Method

Fig. 1 illustrates the proposed method, which consists of two training stages and an inference process.

A. Training of the Bar-wise Feature Extractor and the Bar-wise Composer

We adopt a VAE-based approach [24] to encode each bar x into two latent vectors: one for chord progression (z_{chd}) and the other for texture (z_{txt}). The objective function is:

$$\begin{aligned}&\mathcal{L}_{\text{VAE}} = \\ &-\mathbb{E}_{\substack{z_{\text{chd}}\sim q_\phi \\ z_{\text{txt}}\sim q_\psi}}[\log p_\rho(c|z_{\text{chd}}) + \log p_\theta(x|z_{\text{chd}}, z_{\text{txt}})] \\ &+\text{KL}(q_\phi(z_{\text{chd}}|c)||p(z_{\text{chd}})) \\ &+\text{KL}(q_\psi(z_{\text{txt}}|x)||p(z_{\text{txt}}))\end{aligned} \tag{1}$$

where c is the chord information extracted by a chord extractor f_{chd}, defined as $c = f_{\text{chd}}(x)$. The chord encoder (q_ϕ) uses a bi-directional GRU, and the texture encoder (q_ψ) employs convolutional and recurrent layers to obtain a chord-invariant representation. The chord decoder (p_ρ) reconstructs c from

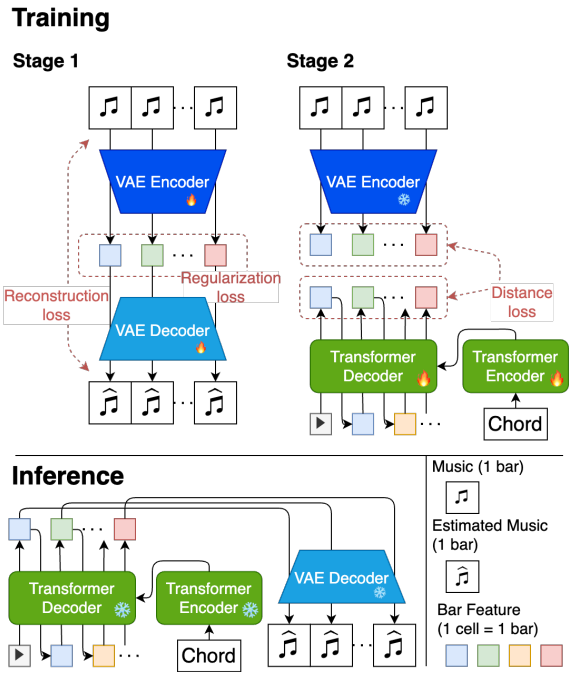

Fig. 1. An overview of the proposed method.

z_{chd}, and the bar-wise composer (p_θ) decodes $(z_{\text{chd}}, z_{\text{txt}})$ into PianoTree-format music [25] using GRU based decoder.

The bar-wise feature vector $\mathbf{bf}_i$ is formed by concatenating z_{chd} and z_{txt} for each bar x_i.

B. Training of the Bar-wise Feature Sequence Generator

Given a sequence of bar-wise features $\mathbf{BF} = [\mathbf{bf}_1, ..., \mathbf{bf}_n]$ and corresponding chord progression $c_s = [c_1, ..., c_n]$, a Transformer-based autoregressive model is trained to model the conditional distribution:

$$p(\mathbf{BF}|c_s) = \prod_{i=1}^{n} p(\mathbf{bf}_i|\mathbf{bf}_{<i}, c_s) \tag{2}$$

where $p(\mathbf{bf}_i|\mathbf{bf}_{<i}, c_s)$ is assumed to follow a normal distribution.

To allow stochastic generation, a vector from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is used as the BOS (Beginning of Sequence) token in place of $\mathbf{bf}_0$. Furthermore, two variants are implemented:

- **mean**: Predicts only the mean of the Gaussian distribution. L2 loss is used for training.
- **gaus**: Predicts both mean and variance, trained via KL divergence.

Although the mean model is deterministic per step, the randomly sampled BOS token induces variation in the generated sequences.

The model is trained with teacher forcing, using ground-truth feature sequences extracted by the trained VAE encoder as input to predict the next. The loss is computed as:

$$\mathcal{L}_{\text{seq}} = \sum_{i=1}^{n} \text{dist}(\mathbf{bf}_i, \hat{\mathbf{bf}}_i) \tag{3}$$

where $\hat{\mathbf{bf}}_i$ is the predicted feature at position i and $\mathrm{dist}(\cdot)$ is either L2 loss (mean) or KL divergence (gaus).

C. Inference Process

During inference, music is generated in three steps:

1) **Bar-wise feature sequence generation**: A sequence $\mathbf{BF} = [\mathbf{bf}_1, ..., \mathbf{bf}_n]$ is generated autoregressively using the bar-wise feature sequence generator, conditioned on a given chord progression. A randomly sampled vector from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is used as the BOS token to introduce stochasticity. In the **gaus** model, each $\mathbf{bf}_i$ is sampled from the predicted Gaussian distribution, further introducing randomness at every step.
2) **Bar generation**: Each feature $\mathbf{bf}_i$ is passed to the VAE decoder to compose the corresponding symbolic segment x_i.
3) **Assembly**: The bars $\{x_1, ..., x_n\}$ are concatenated to form the final composition.

This process allows the model to reflect global structure via the feature sequence while preserving local structure within each bar.

IV. Proposed Metric: BFSD (Bar-wise Feature Similarity Distance)

We introduce **BFSD (Bar-wise Feature Similarity Distance)** as a novel objective metric for evaluating the structural consistency of generated music. BFSD is designed to capture the distribution of similarity patterns among bar-wise feature segments, thereby assessing whether the global structure of the generated music aligns with that of human-composed music.

Let $\mathbf{S}_{\mathrm{ref}} = \{S_{\mathrm{ref}1}, \ldots, S_{\mathrm{ref}n}\}$ be a set of human-composed (reference) pieces, and $\mathbf{S}_{\mathrm{tgt}} = \{S_{\mathrm{tgt}1}, \ldots, S_{\mathrm{tgt}n}\}$ be a set of generated (target) pieces. Each piece S is processed by the bar-wise feature extractor, producing a sequence of bar-wise feature vectors:

$$\mathbf{Z} = [\mathbf{bf}_1, \ldots, \mathbf{bf}_m], \quad \mathbf{bf}_i \in \mathbb{R}^d.$$

To capture repetition and variation patterns over short time spans, we consider subsequences of k consecutive bars and for each subsequence, we concatenate bar-wise feature vectors along the feature dimension to form a combined feature vector:

$$\mathbf{v}_k^{(j)} = [\mathbf{bf}_j^\top \mathbf{bf}_{j+1}^\top \cdots \mathbf{bf}_{j+k-1}^\top]^\top.$$

We then compute pairwise cosine similarities between all such vectors:

$$\mathrm{sim}(\mathbf{v}_k^{(j)}, \mathbf{v}_k^{(l)}) = \frac{\langle \mathbf{v}_k^{(j)}, \mathbf{v}_k^{(l)} \rangle}{\|\mathbf{v}_k^{(j)}\| \, \|\mathbf{v}_k^{(l)}\|}, \quad j \neq l.$$

All similarity values are aggregated over all pieces in the reference and target sets to form empirical distributions. These distributions are converted into histograms with 20 bins, and kernel density estimation (KDE) is applied to obtain smooth distributions p_{ref} and p_{tgt}. The BFSD score for window size k is defined as the KL divergence between these distributions:

$$\mathrm{BFSD}_k = \mathrm{KL}(p_{\mathrm{ref}} \,\|\, p_{\mathrm{tgt}}).$$

Condition	BFSD	BFSD_1	BFSD_2	BFSD_3	BFSD_4
Human	**2.40**	**2.36**	**2.19**	**2.42**	**2.63**
Sample	2.74	2.37	2.60	2.88	3.12
Intra	4.19	4.28	3.92	4.13	4.45
Inter	34.05	27.03	32.58	36.80	39.80
Repeat	368.08	126.68	140.45	598.48	606.70

TABLE I
VERIFICATION RESULTS OF BFSD UNDER VARIOUS STRUCTURAL DISRUPTIONS

The final BFSD score is computed by averaging over window sizes $k = 1, 2, 3, 4$:

$$\mathrm{BFSD} = \frac{1}{4} \sum_{k=1}^{4} \mathrm{BFSD}_k.$$

Lower BFSD values indicate that the generated similarity structure more closely matches that of human-composed music, implying better long-term structural coherence.

V. Experimental Evaluation

A. Verification of BFSD as a Global Structure Metric

To validate the effectiveness of BFSD in evaluating global musical structure, we analyzed its behavior under various types of structural disruptions. Specifically, we computed BFSD between human-composed music and several altered versions of it, each with different levels of structure degradation.

The following sample sets were prepared:

- **Human**: Original human-composed pieces.
- **Sample**: Bar sequences obtained by randomly shuffling bars within 8-bar segments.
- **Intra**: Bar sequences obtained by randomly shuffling bars within the same piece.
- **Inter**: Bar sequences obtained by randomly shuffling bars across different pieces.
- **Repeat**: Sequences where randomly selected bars were repeated to construct 8-bar segments.

We hypothesized that *Sample*, *Intra*, and *Inter* would show decreasing inter-bar relatedness in that order, leading to higher BFSD scores. *Repeat* was expected to unnaturally produce highly similar bars, resulting in exceptionally high BFSD.

For each condition, 160 eight-bar samples were extracted from the POP909 dataset [26].

As shown in Table I, the results support our hypothesis: BFSD increased from *Human* to *Sample*, *Intra*, and *Inter*, while *Repeat* showed the highest score due to excessive similarity. This trend appears consistently across BFSD and all BFSD_k scores ($k = 1$–4).

Notably, BFSD_1 shows little difference between *Human* and *Sample*, since it ignores bar ordering and mainly reflects feature distribution. In contrast, BFSD_2–BFSD_4 show larger differences, confirming their sensitivity to structural coherence. Therefore, averaging over $k = 1$ to 4 provides a reliable definition of BFSD as a global structure evaluation metric.

B. Experimental Setup

For training data, we used the POP909 dataset [26], which consists of 909 professionally arranged solo piano performances of popular songs. From this dataset, we selected 866 pieces that are in either 4/4 or 2/4 time signatures and split them into 90% for training, 5% for validation, and 5% for testing. Each piece was transposed into 12 different keys ranging from -5 to $+6$ semitones for data augmentation.

The hyperparameters for training were set as follows: for the VAE, we set $\beta = 0.1$ and used the Adam optimizer [27] with a learning rate of 1.0×10^{-3} for 100 epochs. The architecture of the VAE components follows [24], with the number of bars per segment set to 1. For the bar-wise feature sequence generator, we used AdamW [28] with a learning rate of 4.0×10^{-4} for 40 epochs. The bar-wise feature sequence generator architecture was based on Llama3 [29], with 8 layers, an embedding dimension of 1024, 4 attention heads, and a RoPE position embedding [30] with a base of 10000.

To evaluate the effectiveness of using a sequence model for bar-wise feature generation, we implemented two baseline models where each bar-wise feature is generated independently using $p(\mathbf{bf}_i|c_{i-k}, ..., c_i, ..., c_{i+k})$ with $k = 0$ or 1. Additionally, we compared our approach with **Polyffusion** [17], a diffusion-based model that generates fixed-length note sequences in a piano roll format. Polyffusion was trained to generate 8-bar-long sequences conditioned on chord information. For generating longer pieces, we used an iterative strategy: the first 8 bars were generated, and the last 4 bars were used as input to generate the next 4 bars.

For each method, we generated 160 pieces in total, consisting of 8-bar and 32-bar compositions using the chord progressions from the test set as input.

C. Objective Evaluation

We evaluated the quality and controllability of generated music using both established objective metrics and a newly proposed metric designed to capture global structure.

We used $\mathbf{D_P}$ and $\mathbf{D_D}$ [31] to evaluate the quality of pitch and duration distributions, respectively. These metrics are defined as the *overlapped area* between the probability density functions (PDFs) of the pitch or duration distributions in the generated music and those in the human-composed reference music. Formally, the score is given by:

$$\mathrm{D_X} = \int \min(p_{\mathrm{ref}}(x), p_{\mathrm{gen}}(x))\, dx, \quad X \in \{\mathrm{P}, \mathrm{D}\}.$$

where $p_{\mathrm{ref}}(x)$ and $p_{\mathrm{gen}}(x)$ are the PDFs of the reference and generated distributions. A higher value indicates better alignment between the distributions.

To evaluate chord controllability, we used **Chord Accuracy (CA)** [31], which measures whether the chords in the generated music match the conditioning chord sequence. Chord Accuracy is defined as:

$$\mathrm{CA} = \frac{1}{N_{\mathrm{tracks}} \cdot N_{\mathrm{chords}}} \sum_{i=1}^{N_{\mathrm{tracks}}} \sum_{j=1}^{N_{\mathrm{chords}}} \mathbb{I}\{C_{i,j} = \hat{C}_{i,j}\},$$

where $C_{i,j}$ denotes the ground-truth chord label at bar j in track i, and $\hat{C}_{i,j}$ denotes the chord estimated from the generated music. A higher CA indicates better harmony control and alignment with the conditioning signal.

In addition, we used **BFSD** to assess the global structure of generated music.

TABLE II
OBJECTIVE EVALUATION RESULTS. THE UPPER BLOCK REPRESENTS 8-BAR GENERATION, WHILE THE LOWER BLOCK REPRESENTS 32-BAR GENERATION.

Method	$\mathbf{D_P}\uparrow$	$\mathbf{D_D}\uparrow$	$\mathbf{CA}\downarrow$	$\mathbf{BFSD}\downarrow$
Polyffusion [17]	**0.896**	0.865	1.148	**3.33**
Baseline $k = 0$	0.631	0.743	2.546	18.03
Baseline $k = 1$	0.538	0.732	2.572	8.39
Proposed mean	0.877	0.874	**0.716**	3.43
Proposed gaus	0.872	**0.877**	1.230	4.97
Polyffusion [17]	0.853	**0.943**	1.67	4.23
Proposed mean	0.802	0.802	**0.717**	2.39
Proposed gaus	**0.898**	0.827	1.950	**1.95**

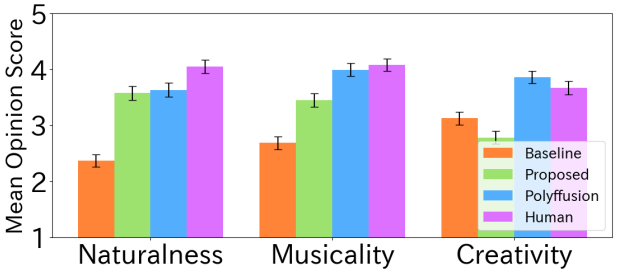

Fig. 2. Subjective evaluation results for 8-bar generation. Error bars indicate 95% confidence intervals.

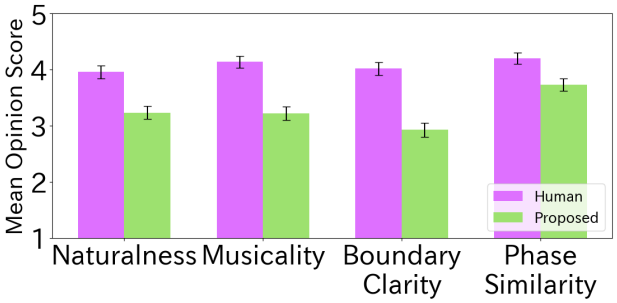

Fig. 3. Subjective evaluation results for 32-bar generation. Error bars indicate 95% confidence intervals.

1) Results: Table II presents the results. For 8-bar generation, the proposed method outperformed the baseline models in all metrics, demonstrating the effectiveness of sequence modeling. Compared to Polyffusion, the proposed method achieved better $\mathrm{D_D}$ and CA, while $\mathrm{D_P}$ and BFSD were slightly worse. For 32-bar generation, the proposed method significantly improved BFSD compared to Polyffusion, indicating its ability to generate long-term structures more effectively. Interestingly, BFSD improved significantly more in 32-bar compositions compared to 8-bar compositions, suggesting a stronger advantage of the proposed method in modeling long-term coherence.

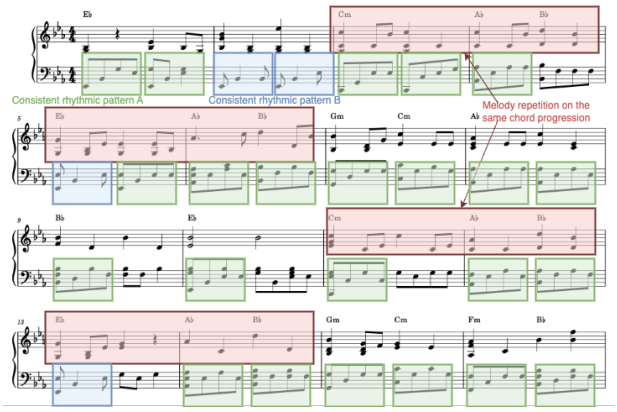

Fig. 4. Excerpt from a 32-bar generated piece (bars 17-32).

D. Subjective Evaluation

For the subjective evaluation, we conducted a five-point MOS (Mean Opinion Score) test with 68 participants (31 with musical experience, 38 without). A participant was considered musically experienced if they met at least one of the following criteria: 1) majored in music or related fields, 2) had at least 3 years of piano performance experience, 3) had at least 3 years of composition or arrangement experience.

For **8-bar generation**, participants listened to four samples each from the baseline model ($k = 1$), the proposed mean model, Polyffusion, and human-composed pieces. They rated them on a five-point scale based on **naturalness, musicality, and creativity**. For **32-bar generation**, participants listened to four samples each from the proposed mean model and human-composed pieces. They rated them on **naturalness, musicality, clarity of section boundaries, and consistency within sections**, again using a five-point scale (higher scores indicate better quality) in the same settings as [23]. Each participant was presented with one of two sets of randomly selected samples.

Fig. 2 and 3 show the results. For **8-bar generation**, the proposed method showed significantly better scores than the baselines in naturalness and musicality, and no significant difference from Polyffusion in naturalness. In comparison to Polyffusion, however, it scored significantly lower in musicality and creativity, suggesting room for improvement. For **32-bar generation**, the proposed method scored significantly lower than human compositions in all criteria, indicating that further refinements are needed for long-term structure generation.

E. Generated Sample Analysis

Fig. 4 shows bars 17-32 from a 32-bar piece generated by the proposed method (mean variant), randomly selected from the test set. Annotations highlight two key structural features. First, the accompaniment part mainly consists of two rhythm patterns, used consistently across bars, contributing to a stable texture. Second, melodic repetition occurs in sections with the same chord progression, indicating that the model can capture global repetition structures.

However, the repeated melodies are overly simple, causing the music to sound monotonous. This suggests that the model tends to generate patterns that are structurally coherent but lacking in expressive variation. These observations are consistent with the subjective evaluation results, where the proposed method received lower scores in creativity and musicality. Together, they indicate a tendency for the model to generate musically coherent but stylistically average compositions.

VI. Conclusion and Future Work

This paper proposed a hierarchical symbolic music generation method based on autoregressive modeling of VAE-based bar-wise features. By representing symbolic music as sequences of bar-level features, the proposed framework aims to efficiently model long-term musical structure while maintaining local coherence.

Objective and subjective evaluations confirmed that the proposed method effectively captures both local naturalness and global structural coherence. However, both subjective evaluation and sample analysis revealed that the generated music often lacks expressive variation. In particular, repeated melodic patterns tend to be overly simple, resulting in musically coherent but stylistically average outputs.

Addressing this issue remains a key challenge. Future work includes enhancing the expressiveness of the bar-wise decoder, refining the feature extraction process to better capture melodic nuances, and incorporating mechanisms to promote diversity and variation across repeated sections.

Acknowledgment

This work was partly supported by JST, CREST, JPMJCR19A3, and JST AIP Acceleration Research JPMJCR25U5, Japan.

References

[1] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proc. of the 28th ACM Int. Conf. on Multimedia*, Seattle, WA, USA: ACM, 2020, pp. 1180–1188. DOI: 10.1145/3394171.3413671. [Online]. Available: https://doi.org/10.1145/3394171.3413671.

[2] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, *Deep Learning Techniques for Music Generation*. Springer, 2019, p. 284.

[3] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press, 2006.

[4] M. T. Pearce and G. A. Wiggins, "Auditory expectation: The information dynamics of music perception and cognition," *Topics in Cognitive Science*, vol. 4, no. 4, pp. 625–652, 2012.

[5] S. Dai, H. Zhang, and R. B. Dannenberg, "Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music," *arXiv preprint*, vol. arXiv:2010.07518, 2020. [Online]. Available: https://arxiv.org/abs/2010.07518.

[6] D. Temperley, *The cognition of basic musical structures*. MIT press, 2004.

[7] F. Lerdahl and R. S. Jackendoff, *A Generative Theory of Tonal Music, reissue, with a new preface*. MIT press, 1996.

[8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[9] C. A. Huang, A. Vaswani, J. Uszkoreit, *et al.*, "Music transformer: Generating music with long-term structure," in *Proc. of the 19th Int. Society for Music Information Retrieval Conf.*, Paris, France, 2018, pp. 726–733.

[10] C. Payne, *Musenet*, OpenAI blog post, 2019. [Online]. Available: https://openai.com/blog/musenet.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint*, vol. 1312.6114, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[13] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[15] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 4361–4370. [Online]. Available: https://proceedings.mlr.press/v80/roberts18a.html.

[16] K. W. Cheuk, R. Sawata, T. Uesaka, *et al.*, "Diffroll: Diffusion-based generative music transcription with unsupervised pretraining capability," *arXiv preprint arXiv:2210.05148*, 2022.

[17] L. Min, J. Jiang, G. Xia, and J. Zhao, "Polyffusion: A diffusion model for polyphonic score generation with internal and external controls," in *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[18] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021, pp. 486–475. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000058.pdf.

[19] 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.

[20] H.-W. Dong and Y.-H. Yang, "Convolutional generative adversarial networks with binary neurons for polyphonic music generation," in *Proc. of the 19th Int. Society for Music Information Retrieval Conf.*, 2018, pp. 190–196.

[21] S.-L. Wu and Y.-H. Yang, "Musemorphose: Full-song and fine-grained piano music style transfer with one transformer vae," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1953–1967, 2023. DOI: 10.1109/TASLP.2023.3270726. [Online]. Available: https://doi.org/10.1109/TASLP.2023.3270726.

[22] B. Yu, P. Lu, R. Wang, *et al.*, "Museformer: Transformer with fine- and coarse-grained attention for music generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 1376–1388.

[23] Z. Wang, L. Min, and G. Xia, "Whole-song hierarchical generation of symbolic music using cascaded diffusion models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=sn7CYWyavh.

[24] Z. Wang, D. Wang, Y. Zhang, and G. Xia, "Learning interpretable representation for controllable polyphonic music generation," in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.*, Montreal, Canada: ISMIR, 2020, pp. 662–669.

[25] Z. Wang, Y. Zhang, Y. Zhang, *et al.*, "PIANOTREE VAE: structured representation learning for polyphonic music," in *Proceedings of the 21th International Society for Music Information Retrieval Conference*, J. Cumming, J. H. Lee, B. McFee, *et al.*, Eds., 2020, pp. 368–375. [Online]. Available: http://archives.ismir.net/ismir2020/paper/000096.pdf.

[26] Z. Wang, K. Chen, J. Jiang, *et al.*, "Pop909: A pop-song dataset for music arrangement generation," *arXiv preprint*, vol. arXiv:2008.07142, 2020. [Online]. Available: https://arxiv.org/abs/2008.07142.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, vol. arXiv:1412.6980, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint*, vol. arXiv:1711.05101, 2019. [Online]. Available: https://arxiv.org/abs/1711.05101.

[29] A. Dubey, A. Jauhri, A. Pandey, *et al.*, "The llama 3 herd of models," *arXiv preprint*, vol. arXiv:2401.00120, 2024. [Online]. Available: https://arxiv.org/abs/2401.00120.

[30] J. Su, A. Murtadha, Y. Lu, S. Pan, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, no. C, 12 pages, Feb. 2024, ISSN: 0925-2312.

[31] Y. Ren, J. He, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Popmag: Pop music accompaniment generation," in *Proceedings of the 28th ACM International Conference on Multimedia*, ACM, 2020, pp. 1198–1206.