Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HarmonyRFG: Bridging Music Theory and Deep Learning through Spiking Neural Networks for Innovative Chord Progression Generation

Zimo Dong

2411060@xdsisu.edu.cn

XianDa College of Economics and Humanities Shanghai International Studies University

Research Article

Keywords: Chord Generation,Transformer,LSTM,SNN,Deep Learning, Real-Time Music Generation

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8757335/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

HarmonyRFG: Bridging Music Theory and Deep Learning through Spiking Neural Networks for Innovative Chord Progression Generation

Zimo Dong

XianDa College of Economics and Humanities, Shanghai International Studies University, 999 Dongtan Avenue, Shanghai, 202150, Shanghai, China.

Contributing authors: dongzimo.work@gmail.com;

Abstract

This research proposes a new music generation model Harmony Rule-based Forward Generation that integrates traditional harmony rules and deep learning. The model combines the global modeling capability of Transformer with the temporal perception characteristics of Spiking Neural Networks, effectively improving the artistry and logic of generated chord sequences. By introducing Markov chains and music theory rules, it achieves automatic generation of chord sequences that conform to human intuitive perception and possess innovation, and in the meantime, showcases the application potential in real-time composition, interactive music installations, human-computer collaboration, and other artistic scenarios.

Keywords: Chord Generation,Transformer,LSTM,SNN,Deep Learning, Real-Time Music Generation

1 Introduction

Music theory provides a structured foundation for organizing harmonic relationships, from classical tonal harmony to chord syntax in popular music. However, many recent data-driven music generation models treat music primarily as discrete symbolic sequences, often lacking explicit theoretical constraints. As a result, generated outputs may exhibit harmonic incoherence and violations of tonal logic. In contrast, traditional rule-based systems preserve harmonic consistency but generally lack the

flexibility, adaptability, and expressive range required in open-ended creative contexts. To address this gap, this research proposes HarmonyRFG (Harmony Rule-based Forward Generation), a music generation framework designed for real-time performance and ambient music creation. By integrating the temporal dynamics of Spiking Neural Networks (SNN) with the global contextual modeling capability of Transformer, and incorporating Markov chains and harmonic rules as structural constraints, HarmonyRFG achieves both musical coherence and generative flexibility. Compared with approaches such as DeepMind's Music RL [1], which often relies on manual reward shaping, and DeepBach [2], which is limited to fixed stylistic domains, HarmonyRFG supports adjustable parameter ranges that enable rapid switching across emotional states, modes, and rhythmic patterns. This allows artists to focus on expressive intent and real-time interaction, expanding the applicability of AI-driven music generation to interactive installations, live performance, and ambient sound environments.

2 Background

Although deep learning has advanced music generation, outputs from many current models are still perceived as lacking artistic depth and emotional expressiveness. A key difficulty is that musical emotion and structure unfold over time, yet many models struggle to represent long-range temporal organization and gradual expressive change. Early studies typically relied on Recurrent Neural Networks (RNNs). Mozer [3] pioneered RNN-based melody generation by integrating pitch, duration, harmonic context, and psychoacoustic features. Eck and Schmidhuber [4] introduced long short-term memory (LSTM) for generating and improvising Blues-style chord progressions, and Boulanger-Lewandowski et al.[5] combined RNNs with Restricted Boltzmann Machines (RBM) to model polyphonic piano-roll representations. However, RNN/LSTM-based approaches often fail to capture long-term musical form, and can collapse into repetitive loop patterns with limited dynamics and emotional variation. More recent attention-based models, such as Pop Music Transformer[6] improve the learning of longer-term structure and chord-progression coherence, but still tend to fall short in rhythmic vitality and fine-grained emotional expression, limiting applicability in scenarios requiring real-time interaction and affective feedback. In parallel, rule-based approaches such as DeepBach combine neural networks with constraint satisfaction to enforce strict stylistic rules for four-part chorales, but remain restricted to specific styles and offer limited flexibility for exploring new expressive spaces. Reinforcement-learning approaches (e.g., Music RL) can enhance diversity and creativity via reward design, yet often require substantial manual intervention and are difficult to deploy for real-time generation and interactive art. To address these limitations in artistry, real-time performance, and style flexibility, this paper proposes HarmonyRFG, a framework designed to generate smooth and musically logical chord sequences in real time while enabling parameter-driven switching across styles and moods. This supports creation workflows where artists prioritize immediate emotional expression, and targets applications such as ambient music, improvisational accompaniment, interactive installations, and music therapy. HarmonyRFG is built on the assumption that the rhythmic pattern of chord changes is as important as harmonic

content for musical style. Accordingly, we combine Transformer with SNN [7] to capture temporal dynamics by jointly learning chord and duration sequences. Chords are thus modeled not merely as discrete symbols but as long-sequence data reflecting evolving musical affect. In addition, Markov transition probabilities and chord-transition scoring rules are introduced as constraints, balancing theoretical plausibility with artistic flexibility and smoothness.

3 Methodology of HarmonyRFG

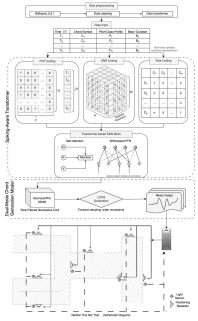


Fig. 1 Operating Mechanism of HarmonyRFG

The automatic music generation framework of HarmonyRFG we propose consists of two core parts (Fig.1):

<1> Spiking-Aware Transformer (SAT)

<2> Dual-Mode Chord Generation Model (DMCG)

We use the billboard_2.0.1 dataset from the McGill Billboard Project for model training, and adopt the proposed syntax for symbolic chord annotations [8, 9].

SAT is based on a hybrid SNN–Transformer architecture. The Transformer captures global structural relationships, while the Feed-Forward Network (FFN) is replaced by an SNN module to model temporal dynamics and duration-sensitive features.

DMCG employs a computationally efficient LSTM decoder for real-time forward chord generation. Chord transition probabilities are provided by the SNN, allowing the LSTM to perform direct forward sampling during inference. Meanwhile, the SNN spike density controls the rhythmic structure of the generated chord sequence, enabling the LSTM to select the next chord and its duration without backtracking. This design ensures fluent chord transitions under low computational cost, improving real-time performance.

In conventional Transformer architectures, the FFN in the Multi-Layer Perceptron (MLP) only transforms instantaneous numerical features and discards temporal dependencies across chords (Fig. 2a, Fig. 2b). As a result, chord duration and long-range temporal associations cannot be effectively modeled. However, musical perception is inherently continuous and holistic. From a Gestalt perspective, listeners tend to perceive music as an integrated temporal structure rather than isolated events [10]. To

capture rhythmic duration and temporal continuity in chord progressions, we replace the FFN with an SNN module.

In the SNN, the neuron's membrane potential (purple membrane potential) accumulates input over time, and a neuron is activated only when the potential reaches a threshold, producing a red spike that triggers a chord transition (Fig. 2c). If a chord needs to be sustained for a longer duration, the neuron fires after a longer interval rather than switching immediately. The resulting discrete spikes, visualized as green bars, encode chord timing, where spike intervals represent chord persistence (Fig. 2d). For example, lower spike density corresponds to longer chord durations (coordinates 2–6), while higher spike density indicates faster chord transitions (coordinates 7–8). Through this spike-based temporal encoding, the model captures rhythmic variation and expressive timing, enabling chord sequences with greater vitality rather than rigid symbolic patterns.

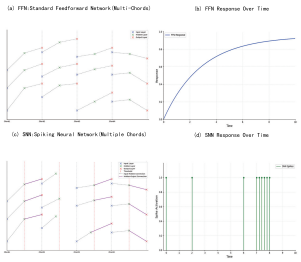


Fig. 2 FFN and SNN Training Method Comparison Chart

Together with this, the Transformer is a neural network architecture based on the self-attention mechanism. Through the Query-Key-Value (QKV) mechanism, it allows each time step in the sequence to be associated with global chord information, simultaneously capturing local and global patterns at multiple levels. This unique structural combination can learn large-span chord structures, tonality regressions, repetitive themes or variations and other chord transition methods in music. It generates more coherent and logical chord sequences, avoiding the frequent fragmentation and jumpiness in traditional models.

We convert the chords in the dataset into fixed-dimensional mathematical representations using the Pitch Class Profile Vector (PCP) method, arranging the pitch characteristics of all chords in matrix form to measure the similarity of chords in pitch distribution, serving as a reference for capturing the transition relationships between chords. Simultaneously, we store the duration of each chord in a matrix to ensure the generated chords have rhythmic changes. Subsequently, this series of information is transformed into pulse signals, allowing SNN to directly accept temporal information, enabling it to learn both the methods of chord transitions and the rhythms of chord progressions.

In addition, to ensure the generated chord sequences possess both artistic innovation and musical theoretical reasonableness, we introduce Markov transition probabilities and chord transition scoring rules into the attention mechanism of Transformer. Markov transition probabilities can constrain unreasonable chord jumps, ensuring the

model can effectively control the fluency of chords during training, while music theory rules endow the model with harmonic logic and color changes, making the generated music more in line with auditory habits and aesthetic preferences. Secondly, we combine the relative frequencies of chord categories in the dataset, encode music rules and set different weights, enabling the model to more accurately identify the transition patterns of functional harmony and color harmony, thereby enhancing the diversity and expressiveness of chord colors while maintaining the reasonableness of music theory.

4 Results and Analysis

To evaluate the performance of HarmonyRFG, we analyze the model from two complementary perspectives: technical performance and artistic expression. Technical evaluation focuses on training loss and computational energy consumption, reflecting efficiency and real-time feasibility. Artistic evaluation examines auditory perception of the generated music, assessing musical logic, expressiveness, and overall musicality.

Regarding training efficiency and energy consumption, we compare the loss convergence of SNN and FFN during training (Fig. 3a, Fig. 3b). The FFN shows a slower loss decline and converges to a higher final loss, whereas the SNN converges more rapidly with a lower final discrepancy. The SNN reaches convergence approximately 50 epochs earlier, indicating more effective learning of temporal dynamics in chord sequences. Moreover, SNNs perform computation only upon neuron spiking, while FFNs process all inputs at each step. This event-driven mechanism reduces model complexity by approximately 40%, significantly improving computational efficiency and making the model well suited for real-time generation and low-power deployment scenarios such as neuromorphic hardware and embedded systems.

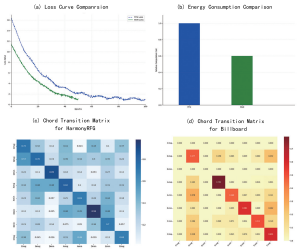


Fig. 3 Model Training Results and Evaluation

In the results of model training, we visualized the chord transition probability matrix of a randomly selected set of music generated by HarmonyRFG using a heatmap (Fig.3c). The horizontal and vertical axes represent different chord types, and the color is darker along the diagonal, showing an obvious weighting trend. For example, Dmin -> Dmin = 0.98, indicating that the probability of chords remaining unchanged is high, and chord progressions repeating this transition pattern are more common, which conforms to the regularity of chord stability in music and reflects the generated music having a certain sense of stability and continuity. In contrast, rare chord jumps such as Cmaj -> Emin = 0.13 have a lower probability of occurrence, indicating that

the model tends to follow musical logic during the learning process, using smooth chord transitions rather than drastic jumps.

Subsequently, we compared this set of chord transition probabilities with the chord transition probabilities in the original dataset (Fig.3d). In the difference of diagonal elements: the blue matrix considers the persistence of chords, while the yellow matrix hardly does. This indicates that the trained model has learned the characteristic that chords can persist. In the non-diagonal elements, although some transition probabilities are relatively close between the two, such as Gmaj -¿ Cmaj, there are also some significant differences, such as Cmaj -> Gmaj, Amaj -> Amin, Dmin -> Cmaj, Emin -> Gmaj. This means that the trained model has a deviation from the original dataset in the probabilities of some chord transitions. The trained model will try to learn the general patterns in music rather than reproducing the probability distribution of the original dataset, and may generalize some transition patterns that are not so obvious in the original dataset. The heatmap of HarmonyRFG shows that the model learns more different chord functions, and the transition probabilities between functional chords are often higher than those between non-functional chords, which conforms to the theory of traditional harmony.

Finally, we compared the chord duration distributions in the original dataset and the generated dataset (Table.1). Overall, the difference between the data generated using the HarmonyRFG model and the original dataset is small. Among them, the values of the average duration are very close, indicating that the chord lengths generated by the model are close to the real data. The average chord duration of HarmonyRFG is slightly longer, its chords are more stable, change more slowly, and tend to have longer durations. From the comparison of standard deviations and minimum values, due to the higher weights given to more stable chord sequences during training, the chord duration fluctuations of the model are smaller, slightly less than the original data, and biased towards generating more stable chord sequences. Simultaneously, due to the influence of rule constraints, the probability of generating extremely short chords is reduced.

Table 1 The Chord Duration Distribution of HarmonyRFG Generated Data and Billboard Data

	HarmonyRFG data	McGill-Billboard data
Average duration	1.590 seconds1	1.542 seconds
Median duration	1.718 seconds	1.531 seconds
Standard deviation	1.052 seconds	1.273 seconds
Minimum value	0.073 seconds	0.023 seconds

In actual artistic creation, the auditory perception and style changes of the generated music can be adjusted through the two sets of adjustable parameters: the membrane potential threshold and the Markov chain weight. The membrane potential threshold determines when neurons trigger spikes and generate new chords. A lower threshold means neurons are more easily activated, allowing the model to respond

agilely to environmental or interactive mood changes, suitable for interactive art installations and live improvisational music performances. A higher threshold means neurons are more difficult to trigger, making the generated chords more stable, suitable for continuously creating a stable background music atmosphere. In addition, the Markov chain weight determines the strictness of music theory constraints. A higher weight ensures the fluency and theoretical rigor of the generated music, helping to create a stable and harmonious auditory effect in artistic creation (Fig.4a). Conversely, a lower weight allows the model to explore more free, more indeterminate chord progressions with more experimental, artistic expressiveness and richer changes in chord transitions (Fig.4b).

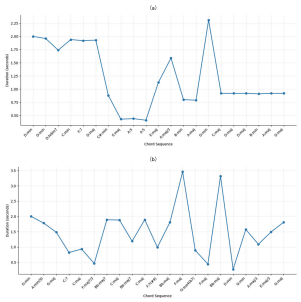


Fig. 4 Comparison of Chord Durations Generated with Different Weights

To further explore the perceptual characteristics of music generated by HarmonyRFG, an informal listening evaluation was conducted based on qualitative auditory feedback. A group of 20 listeners (aged 17–30) were invited to listen to short music clips generated in real time under different parameter settings. In total, eight clips were presented, each corresponding to a distinct combination of membrane potential threshold and Markov chain weight.

Listeners provided perceptual ratings on a five-point scale across three dimensions: music fluency, emotional expression, and innovativeness. The feedback revealed consistent perceptual trends across parameter configurations. Higher Markov chain weight settings were generally associated with smoother chord transitions, greater harmonic stability, and a calmer emotional character, whereas lower weight settings were perceived as more exploratory, exhibiting freer chord progressions and increased rhythmic variation. These observations suggest that HarmonyRFG enables controllable modulation of musical structure and expressiveness through parameter adjustment.

From the listening feedback, configurations with medium–low membrane potential thresholds and medium Markov chain weights tended to achieve a perceptual balance between harmonic coherence and expressive variability. This parameter region was therefore selected as a representative operating range for subsequent application-oriented demonstrations. Based on this setting, the following case study illustrates how HarmonyRFG functions as a parametric generative system within an immersive art healing environment, where model parameters are dynamically influenced by contextual and environmental inputs.

5 Case Study

We presented Neither This Nor That at the Zhujiajiao Ultra-Immersive Art Season in Shanghai, China, which explored the aesthetic relationship between real-time sound generation and immersive spatial experience (Fig. 5a). The work demonstrates the real-time generative capability of HarmonyRFG under predefined procedural constraints, emphasizing a process-oriented approach to immersive music experience rather than predetermined musical outcomes [11]. Trained on a popular music dataset, HarmonyRFG learns common chord progressions and stylistic features. Through parametric control during generation, the system preserves musical familiarity while improving harmonic balance and timbral continuity, resulting in smooth transitions and a stable emotional flow suitable for immersive listening.

HarmonyRFG operates as part of an integrated generative art system responsive to environmental input. Light sensors embedded in the installation capture variations in illumination and time, which are mapped to real-time modulation of musical parameters. Eight spatially distributed speakers (Fig. 5b) enable dynamic allocation of sound across the space, forming a layered soundscape. Changes in light intensity subtly influence local volume, rhythmic density, and musical texture, allowing sound to evolve in close correspondence with spatial lighting conditions. Through this interaction, time, space, music, and light are unified into a coherent experiential system [12].

On-site observation indicated that many audience members remained in the installation space for extended periods, often slowing their movement or closing their eyes. Informal feedback suggested that the fluidity and harmonic balance of the music promoted relaxation and sustained attention, while spatial reverberation enhanced emotional immersion (Fig. 6) [13]. These observations suggest that the real-time, environmentally driven nature of HarmonyRFG supports emotionally resonant and introspective listening experiences within immersive artistic contexts.

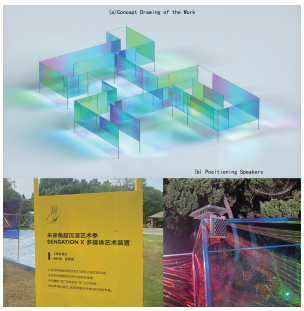


Fig. 5 Zimo Dong, Neither This Nor That, Laser PVC, Aluminum and Positioning Speakers, 13mX9mX2m,2023

6 Discussion and Summary

From a technical perspective, HarmonyRFG improves the coherence, logical consistency, and emotional expressiveness of generated chord sequences by combining



Fig. 6 Zimo Dong, Neither This Nor That, Laser PVC, Aluminum and Positioning Speakers,13mX9mX2m,2023, On-site Audience and Artists Viewing the Works

the global contextual modeling of Transformer with the temporal sensitivity of SNN. Compared with RNN-based or strictly rule-driven methods, the framework demonstrates clear advantages in real-time generation, smooth chord transitions, and adaptive mood expression. By integrating data-driven learning with harmonic constraints, HarmonyRFG provides a flexible and musically grounded approach suitable for applications such as music healing, interactive art, and generative music systems.

Beyond algorithmic performance, HarmonyRFG functions as an interactive component within the creative process. Its parameterized structure enables creators to adjust harmonic organization, rhythmic density, and emotional character in real time in response to contextual conditions. This interactive capability supports improvisational workflows and real-time artistic exploration, extending the expressive potential of AI-assisted music generation. Rather than being evaluated solely by technical metrics, the system's value lies in its capacity to facilitate emotional engagement, audience resonance, and novel auditory experiences. In this sense, HarmonyRFG expands the design space for musical performance, immersive installations, and interdisciplinary sound practices. Future work will further enhance its expressive flexibility and adaptability, supporting broader applications in generative and interactive art contexts.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Cideron, G., et al.: MusicRL: Aligning Music Generation to Human Preferences. https://arxiv.org/abs/2402.04229. arXiv:2402.04229 (2024)

[2] Hadjeres, G., Pachet, F., Nielsen, F.: Deepbach: A steerable model for bach chorales generation. In: Proceedings of the 34th International Conference on Machine Learning (ICML), vol. 70, pp. 1362–1371 (2017)

[3] Mozer, M.C.: Neural network music composition by prediction: Exploring the benefits of psychophysical constraints and multiscale processing. Connection Science

6, 247–280 (1994)

[4] Eck, D., Schmidhuber, J.: Learning the long-term structure of the blues. In: Proceedings of the International Conference on Artificial Neural Networks (ICANN), pp. 284–289 (2002)

[5] Boulanger-Lewandowski, N., Bengio, Y., Vincent, P.: Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In: Proceedings of the 29th International Conference on Machine Learning (ICML), pp. 1881–1888 (2012)

[6] Huang, Y.-S., Yang, Y.-H.: Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1180–1188 (2020)

[7] Yao, M., Qiu, X., Hu, T., Hu, J., Chou, Y., Tian, K.: Scaling spike-driven transformer with efficient spike firing approximation training. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2973–2990 (2025)

[8] Burgoyne, J.A., Wild, J., Fujinaga, I.: An expert ground truth set for audio chord recognition and music analysis. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), pp. 633–638 (2011)

[9] Harte, C.A., Sandler, M.B., Abdallah, S.A., Gómez, E.: Symbolic representation of musical chords: A proposed syntax for text annotations. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR), pp. 66–71 (2005)

[10] Sapaty, P.S.: Towards wholeness and integrity of distributed dynamic systems. Journal of Computer Science & Systems Biology **9**(3) (2016)

[11] Miranda, E.R.: On computational models of musical creativity. Contemporary Music Review **22**(4), 25–47 (2003)

[12] Bérigny, C.D., et al.: Eeg and sonic platforms to enhance mindfulness meditation. Journal of Arts and Humanities (5), 1–12 (2016)

[13] Bro, M., et al.: Live music in a hemodialysis setting: A qualitative study on patient, nurse, and musician perspectives. Healthcare **10** (2022)