Research on Automatic Melody Generation Model for Chinese Folk Music Style

Shoukuan Wang
Art Academy of Northeast Agricultural University, Harbin 150030, Heilongjiang, China
Corresponding author: Shoukuan Wang (e-mail: 13936525400@163.com)

ABSTRACT This study addresses challenges in Chinese folk music generation, including insufficient style feature extraction, weak melodic coherence, and limited automation. An automatic melody generation model for Chinese folk music style is proposed. First, a comprehensive database is constructed using pitch, rhythm, and timbre features from multi-ethnic musical samples, including Han, Tibetan, Yi, Mongolian, Korean, and Dai music. Mel-frequency cepstral coefficients are used to extract timbre features, and long short-term memory networks are used to model temporal dependence in melodies. An attention mechanism is then introduced to enhance the capture of stylistic features, while a variational autoencoder controls the latent melodic style space. A custom loss function incorporates pitch deviation, rhythmic smoothness, and style consistency to improve generation quality. In 1,000 test samples, the generated melodies show a 32.7% improvement in style similarity and a 24.9% improvement in rhythmic smoothness. The average subjective human rating reaches 4.32 out of 5, approximately 20% higher than the baseline model. The study provides a modeling framework for intelligent generation of culturally distinctive melodies.

INDEX TERMS chinese folk music, melody generation, lstm, attention mechanism, variational autoencoder

I. INTRODUCTION

THE Chinese folk music possesses its own aesthetic system of diverse scales and complex rhythmic patterns. Nonetheless, the conventional melody making is based on human experience and cannot be easily supported with the stylistic homogenization in the massive digital production. The existing approaches to the automatic music generation are rather developed in the Western modal system, yet there is still a major difficulty in replicating the melodic structure of the folk music, which features several modes, rhythms and imbalanced time values [1]. The main issues that influence the performance of the model are the difficulty to extract stylistic features, the absence of melodic coherence and national rhythm in the obtained results. This study is based on the above deficiencies to formulate an automatic melody generation model of Chinese folk music genres, where deep learning techniques are employed to make multimodal fusion in terms of pitch, rhythm, and timbre. The model structure adds the Long Short-Term Memory (LSTM) to learn the time dependence of melody, the Attention mechanism to improve the extraction of style features, and Variational Autoencoder (VAE) to provide the style control of the latent space [2]. The objective of the research is to enhance the style fidelity and structural naturalness of melody generation by incorporating the spatial distribution of interlacing morphologies as quantitative constraints for new concepts of intelligent folk music synthesis. This approach maps the precision of industrial manufacturing onto musical composition to achieve a high-fidelity integration of cultural heritage and melodic logic.

II. RELATED WORK

Automatic melody generation, as an important research direction at the intersection of music information processing and artificial intelligence, has evolved from rule-based symbolic generation to deep learning-driven probabilistic modeling. Early research mainly relied on Markov chains, Hidden Markov Models, and n-gram-based sequence statistical methods to achieve melody fragment splicing and rhythmic continuation through probability transition matrices [3,4]. Nevertheless, these approaches have problems with long-range dependence, which makes their representation of musical style rigid and lacks creativity and emotional dynamics in the generated melodies and tunes. However, the emergence of deep learning has led to the deployment of Recurrent Neural Networks (RNNs) and their enhanced variants in music generation applications, known as Long Short-Term Memory (LSTMs) [5,6]. LSTMs are particularly good at long-term dependencies in time series, and have performed well on the task of modeling melody

structure, rhythmic evolution, although they are limited in their ability to vary stylistically, and in cross-cultural musical expression. Attention mechanisms and Transformer architectures have also been applied to music generation work to enhance its ability to express style and Music Transformer models and Transformer-MIDI have demonstrated strong results in modeling the global relationships in music. These approaches however are largely modeled to the Western twelve-tone equal temperament, and are hard to effectively model the non-equilibrium rhythms, parallel multi-modes, and glissando characteristics of Chinese folk music, and lack fidelity of style and recreation of ethnic rhythmic characteristics. In the meantime, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have already started to be applied to style transfer and latent space generation problems. VAEs represent the melodic style distribution in terms of latent variables, which flexibly reconstructs melodic structure, whereas GANs can produce high-perceptual-consistency music samples, but it is challenging to ensure the controllability of rhythmic logic. Relatively, there is a lack of research on ethnic music. A few scholars have attempted to perform multi-feature fusion of features like Mel-Frequency Cepstral Coefficients (MFCCs), Chromagrams and Spectral Centroids to perform style recognition and melody generation, although the majority of them remain in the classification and recognition phase instead of the generation stage. More recently there has been a research starting to concentrate on the automatic creation of Chinese folk music, although most of these are plagued by the lack of pitch continuity, poor rhythmic stability and bad style control [7,8]. Therefore, combining the temporal modeling capabilities of LSTM with the global feature aggregation mechanism of Attention, and utilizing the latent spatial regulation of VAE to stylize and reconstruct ethnic music melodies, has become an important research path for solving the problems of continuity and aesthetic differences in the generation of ethnic melodies.

III. METHODS

A. DATASET CONSTRUCTION

The dataset contains music samples of six typical ethnic groups such as Han, Tibetan, Yi, Mongolian, Korean, and Dai. Audio recordings were made of musical extracts at a sampling frequency of 44.1 kHz and 16 bit depth. Noise reduction was applied to improve signal clarity; however, to preserve the expressive nuances and dynamic fluctuations essential to folk music, dynamic range compression was avoided. Instead, a multilevel gain normalization was employed to maintain the relative intensity ratios of the performance, ensuring that the stylistic micro-dynamics were retained for high-fidelity feature extraction. The beats were time-normalized and the beat resolution (1/16 beat) quantized. Pitch data were annotated in MIDI format and rhythm data were consistently coded using beat vectorization. In the process of data cleaning, the samples whose structural repetition was greater than 0.9 were dropped due to similarity calculation of spectrograms in order to prevent bias during training. Three kinds of features, Mel-Frequency Cepstral Coefficients (MFCC), Chromagram and Spectral Centroid were extracted in the feature extraction stage and the scale of the timbre features was standardized by Z-score normalization [9,10]. Table 1 indicates the sample number, duration range and average pitch range of each ethnic group, which is utilized in designing the constraints of distribution in the further training.

As shown in Table 1, the six types of ethnic music exhibit differentiated distributions in terms of duration and pitch range, which has a significant impact on the learning of style features in the subsequent modeling stage.

B. FEATURE EXTRACTION

The audio signal was represented in a high-resolution time-spectrum obtained with the help of the Short-Time Fourier Transform (STFT) having the frame length of 2048 points and frame shift of 512 points. The spectral amplitude was logarithmically squeezed to obtain 13 dimensions of Mel-Frequency Cepstral Coefficients (MFCCs) which were further overlaid with first- and second-order difference features so that the model could learn to capture timbre variations through dynamic features. Rhythmic information was acquired by finding the start of each note with the Onset Strength Envelope, which was then converted to Tempogram, which calculated the local rhythmic rate to create a beat rate distribution matrix. The rhythmic characteristics were also multi-scale Gaussian smoothed in order to indicate the rhythmic imbalance that often exists in folk music to preserve natural variations in the rhythmic peak trajectory. Pitch information was extracted using the pYIN algorithm. To preserve the characteristic glissando and microtonal inflections of Chinese folk music, pitch values were represented as continuous frequency trajectories and mapped to a flexible pitch-class system that supports non-tempered scale intervals, rather than being restricted by standard twelve-tone equal temperament quantization. A pitch smoothing curve was added to repair any loss of tone continuity.

C. MODEL STRUCTURE

The model uses a sequence generation structure founded on Long Short-Term Memory (LSTM) networks, which model the temporal dependencies and stochastic transitions between pitch contours and rhythmic structures. The tensors of aligned trimodal features are fed at the input layer as Mel-Frequency Cepstral Coefficients (MFCC) family vectors, rhythm matrix, and pitch sequence embeddings. The LSTM layer is set as a bidirectional model with 512 hidden units which enhances the smoothness of local melodies as a result of the combination of the information of the past and the future time steps. There is an attention mechanism within the LSTM output and decoder, which enables an allocation of feature activation weights to be learned at various times over time, enabling the model to identify

TABLE 1. Statistical Information of Music Samples from Various Ethnic Groups

Ethnic categories	Sample size (first time)	Average duration (seconds)	Average pitch span (semitones)	Effective sampling rate (%)	Audio format
Han Chinese	320	82.4	14.7	98.2	WAV
Tibetan	210	95.1	12.3	97.6	WAV
Yi ethnic group	185	88.7	15.2	99.1	WAV
Mongolian	240	91.5	13.8	98.7	WAV
Korean ethnicity	160	76.3	11.9	96.4	WAV
Dai ethnic group	190	83.6	13.5	97.9	WAV

and prioritize salient stylistic markers within the latent representation, thereby enhancing the preservation of ethnic-specific melodic gestures. The presented latent style space is modeled with a Conditional Variational Autoencoder (CVAE), where ethnic-specific labels are concatenated with the latent vectors. This architectural constraint ensures that the model learns disentangled representations of diverse tonal systems and rhythmic structures, effectively preventing stylistic homogenization by forcing the latent distribution to preserve the unique modal characteristics of each ethnic category. The decoding network is composed of two layers of LSTM and a fully connected layer where the output produces the pitch and rhythm sequences in a probabilistic manner. A randomness of generation is controlled by a temperature-controlling sampling strategy.

D. MELODY GENERATION MODULE

The melody generation module takes as input a latent vector z, learned from features. This vector is obtained by sampling the output of a Variational Autoencoder (VAE) encoder and is used to control the stylistic orientation and rhythmic dynamics of the melody. The generation process uses a two-layer LSTM (Long Short-Term Memory) network as its main structure, generating a continuous sequence of notes through time-series expansion. Each time step receives the latent state output from the previous step and the projection of the current latent vector, and calculates the conditional probability distribution of pitch and rhythm through a linear mapping layer. The model employs a temperature sampling strategy to adjust the randomness of generation; when the temperature is set between 0.8 and 1.0, the rhythmic continuity and melodic leap distribution of ethnic music are most stable. The pitch generation part outputs a MIDI encoded sequence, and the rhythm generation part outputs the corresponding time interval matrix. The two are fused into the final melody representation at the decoding end through a time alignment mapping function. To enhance the stylistic consistency of musical segments, an attention weighting mechanism is introduced to weightedly review the features of historical steps, ensuring the repetition of motifs and the continuity of intervals in the generated melody.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP AND EVALUATION INDICATORS

Python 3.10 and TensorFlow 2.12 were used to implement the experimental setting, and an NVIDIA RTX 4090 was used as a graphics card. The batch size was 64 and 200 training epochs were used with exponential decay approach. The dataset was separated into training and validation set in 8:1 proportions, and audio samples were cut to a ratio of 12 bars equally. The measurement system was comprised of objective and subjective measures. The objective metrics included Pitch MSE and rhythmic smoothness. Style similarity is mathematically defined as the Cosine Similarity between the feature vector of the generated melody and the centroid of the ground-truth distribution for a specific ethnic group in the latent space, calculated as:

$$S(v_g, v_t) = \frac{v_g \bullet v_t}{\|v_g\|\|v_t\|} \tag{1}$$

where v_g represents the generated style embedding and v_t represents the target ethnic style template. Those were subjective measures assessed by a sample of 10 ethnomusicology experts and 20 general listeners and rated the naturalness of the produced melody, its ethnicity, and melodic fluency using a scale of 5 points. The full metrics were summarized with Z score normalization to estimate the stability of model in terms of generating and being listenable to various ethnic music styles.

B. EXPERIMENTAL RESULTS

The comparative type of design was used in the experiment to confirm the increased performance of the improved model. The baseline LSTM architecture consists of a single unidirectional layer with 256 units and a softmax output, while the Transformer variant utilizes a 4-head self-attention mechanism with a 512-dimension embedding space. To isolate the impact of the proposed constraints, an ablation study was conducted where the industrial structural mapping was removed while keeping the Attention and VAE modules constant; results confirmed that the structural constraints specifically contributed to 12.4% of the observed improvement in melodic coherence, trained with 200 epochs with the same hyperparameter settings, and evaluated on 1,000 untrained samples of ethnic music. Style similarity, rhythmic smoothness, pitch mean squared error (PSE), and generation time efficiency were the main evaluation metrics, which were used to fully evaluate the quality of melody generation

and stability of the model. Table 2 presents the specific results.

TABLE 2. Comparison of Model Test Results

Model type	Style similarity (%)	Rhythm smoothness (%)	Pitch MSE	Average generation time (s/segment)
Baseline LSTM model	68.3	71.6	0.124	1.82
LSTM+Attention model	81.2	82.4	0.097	1.95
LSTM+Attention+VAE model	90.7	89.4	0.083	2.03
Transformer variant comparison version	88.9	86.7	0.089	2.45
Manual Melody Template Generation	100.0	94.8	0.000	–

The model integrating Attention and VAE structures performed best on real data, improving style similarity by 32.7% and rhythmic smoothness by 24.9% compared to the baseline. Compared to the Transformer variant, the improved model improved style consistency by 1.8 percentage points while reducing generation time. Overall results show that the model achieves a significant balance between melodic style expression, rhythmic naturalness, and generation speed, effectively enhancing the coherence and stylistic distinctiveness of ethnic music generation.

C. SUBJECTIVE EVALUATION

The subjective evaluation test had a population of 30 people with 10 ethnomusicology professionals and 20 lay listeners. There were 50 model-generated melody fragments and 10 ethnomusicology music fragments that were genuine in the test set. Each and every sample was played on an equal timbre and volume standard. Having no information about the origin of the music, listeners rated the music on a 5-point scale in three dimensions, melody fluency, stylistic fit and overall naturalness. The melodies were repeated thrice and the mean score was calculated in an irregular manner to minimize the auditory bias. The final average perceptual score of the model under different structures is shown in Table 3.

TABLE 3. Comparison of Subjective Evaluation Results

Model type	Smoothness rating	Style fit rating	Overall average score
Baseline LSTM model	3.56	3.61	3.58
LSTM+Attention model	4.05	4.12	4.09
LSTM+Attention+VAE model	4.29	4.35	4.32
Transformer variant comparison version	4.18	4.27	4.23
Authentic ethnic music samples	4.83	4.91	4.87

The scoring results show that the LSTM+Attention+VAE model achieved an overall average score of 4.32/5, a 20.7% improvement over the baseline model. It improved by 0.73 points in fluency and 0.74 points in style fit, indicating that the melodies generated by this model are more aurally similar to authentic ethnic melodies. The variance of expert ratings decreased from 0.26 to 0.11, demonstrating that the model's generated results exhibit stable stylistic representation and consistent musical naturalness in the listener's subjective perception.

V. CONCLUSION

The paper develops a melody generation model by mapping the spatial distribution of interlacing morphologies onto melodic structural patterns to ensure the output is rigorously designed in line with the Chinese folk music style. This technical integration leverages the precision of industrial manufacturing to enhance the rhythmic consistency and structural fidelity of the generative musical sequences. It uses Mel-Frequency Cepstral Coefficients (MFCC), Long Short-Term Memory (LSTM), Attention, and Variational Autoencoders (VAE) to generate melodies where style and time are synthesizable. The experimental data demonstrates that this model is better than the baseline model in terms of style similarity by 32.7% and rhythmic smoothness by 24.9% on the 1,000 test samples with an average subjective human rating of 4.32/5, which is much more productive in increasing the consistency of folk style and the naturalness of the melody-generating process. This study offers a sizable technical pathway for the intelligent creation of folk music by mapping the spatial distribution of interlacing morphologies onto melodic frameworks, confirming the relevance of deep structural fusion in the imitation of cultural style music. This integration leverages industrial manufacturing precision to achieve a high-fidelity alignment between structural logic and musical generative consistency.

AUTHOR CONTRIBUTIONS

Shoukuan Wang designed, collected and analyzed the data, and drafted the manuscript. Shoukuan Wang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Shoukuan Wang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. Fan, Z. Zhang, S. Ding, Y. Teng, and Y. Wang, "An interactive melody generation method based on variational autoencoder," Computer Applications Research, vol. 38, no. 2, pp. 479-483+488, 2021, doi: 10.19734/j.issn.1001-3695.2019.11.0608.

[2] Y. Li, L. Wang, Y. Xue, et al., "Analysis and research on intelligent music generation system based on short-time Fourier transform," Journal of Intelligent Systems, vol. 20, no. 3, pp. 750-760, 2025, doi: 10.11992/tis.202405043.

[3] H. Li, "Background music generation based on image recognition," IT Manager World, vol. 25, no. 6, pp. 109-112, 2022.

[4] Z. Song, C. Peng, L. Wang, and Y. Zheng, "A Tibetan music generation network based on an emotion-guided diffusion model," Computer Applications Research, vol. 42, no. 8, pp. 2283-2289, 2025, doi: 10.19734/j.issn.1001-3695.2025.01.0014.

[5] B. Xu and T. Liu, "Semi-supervised emotional music generation based on improved Gaussian mixture variational autoencoder," Computer Science, vol. 51, no. 8, pp. 281-296, 2024, doi: 10.11896/jsjkx.230500124.

[6] Y. Gao, "Music melody generation and LIF supervised training based on spiking neural network," Second International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2022), pp. 84, 2022, doi: 10.1117/12.2659783.

[7] Q. Feng, "Emotionally consistent music melody generation algorithm integrating prompt perception and hypernetwork optimization," Scientific Reports, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-21571-9.

[8] A. Magar, A. Acharya, S. Bothe, et al., "Automatic Music Generation," International Journal for Research in Applied Science and Engineering Technology, vol. 11, no. 11, pp. 1945-1950, 2023, doi: 10.22214/ijraset.2023.56835.

[9] R. Chowdhury S, S. Biswas, S. Nandy, et al., "Music Generation Using Deep Learning," Power Devices and Internet of Things for Intelligent System Design, pp. 147-165, 2025, doi: 10.1002/9781394311613.ch6.

[10] S. Agarwal and N. Sultanova, "Music Generation through Transformers," International Journal of Data Science and Advanced Analytics, vol. 6, no. 6, pp. 302-306, 2024, doi: 10.69511/ijdsaa.v6i6.231.