An Order-Complexity Aesthetic Assessment Model for Aesthetic-aware Music Recommendation

Xin Jin, Wu Zhou, Jinyu Wang
Department of Cyber Security, Beijing
Electronic Science and Technology
Institute, Beijing, 100070, China
Beijing Institute for General Artificial
Intelligence, Beijing, 100085, China
jinxin@besti.edu.cn

Duo Xu*
Tianjin Conservatory of Music,
Tianjin, 300171, China
Qingdao Academy of Intelligent
Industries, Qingdao, 266114, China
Tongji University, Shanghai, 200082,
China, many33@126.com

Yongsen Zheng*
School of Computer Science and
Engineering, Sun Yat-sen University,
No. 132, East Outer Ring Road,
Guangzhou-College-Town,
Guangzhou, 510006, China
z.yongsensmile@gmail.com

ABSTRACT

Computational aesthetic evaluation has made remarkable contribution to visual art works, but its application to music is still rare. Currently, subjective evaluation is still the most effective form of evaluating artistic works. However, subjective evaluation of artistic works will consume a lot of human and material resources. The popular AI generated content (AIGC) tasks nowadays have flooded all industries, and music is no exception. While compared to music produced by humans, AI generated music still sounds mechanical, monotonous, and lacks aesthetic appeal. Due to the lack of music datasets with rating annotations, we have to choose traditional aesthetic equations to objectively measure the beauty of music. In order to improve the quality of AI music generation and further guide computer music production, synthesis, recommendation and other tasks, we use Birkhoff's aesthetic measure to design a aesthetic model, objectively measuring the aesthetic beauty of music, and form a recommendation list according to the aesthetic feeling of music. Experiments show that our objective aesthetic model and recommendation method are effective.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**.

KEYWORDS

Computational aesthetics, Music recommendation, Music evaluation, Birkhoff's measure

*Duo Xu and Yongsen Zheng are corresponding authors.

1 INTRODUCTION

Computational aesthetic evaluation [20, 21] refers to using computational methods to assess the aesthetic beauty of different types of art, including paintings, photographs and music. It usually makes qualitative or quantitative evaluations [32] to assess the quality of artworks generated by AI. It is necessary to consider the aesthetics of artworks because beauty often make people feel pleasant [13].
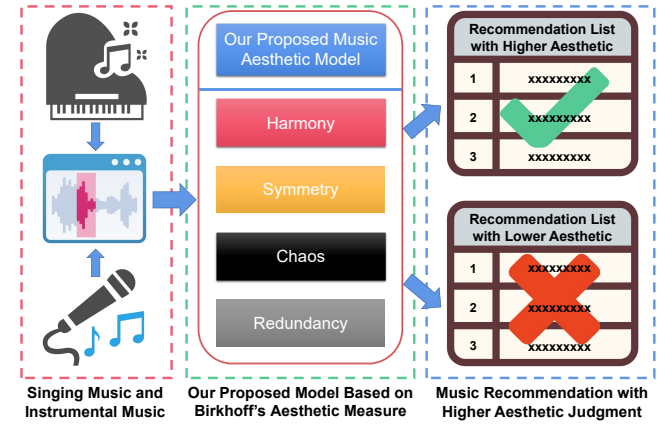

Figure 1: We construct a music recommendation list based on the results of the proposed music aesthetic model by adding four aesthetic features into music recommendation task.

In recent years, recommendation systems have flourished in various fields [62], including music related fields. Nowadays, when faced with such a massive amount of music resources, music recommendation can help users quickly find music of interest [60]. In fact, traditional recommendation systems often recommend music based on user preferences or the popularity of the music [59], which may overlook some music works with higher artistic value.

Generally speaking, the beauty of artistic works, especially the beauty of music, is highly abstract and subjective. The earliest research on music aesthetics is proposed by mathematicians (tracing back to Pythagoras) and attempt to solve it by abstracting it into emotional expression [46] or language [39].

We attempt to propose a method to objectively evaluate the beauty of music. Fig.1 briefly shows our work.

A musical art work usually needs to go through three steps from creation to hearing like Fig.2: Firstly, the musician composes and arranges music to make music score. Secondly, the performer plays

music score to get performance. Finally, rendering the performance into sound according to different instrument selections [51].

However, these music data are not labeled with aesthetic ratings like AVA [48] or AADB [37] in the field of image aesthetic evaluation. So we can't predict the beauty of music due to the lack of subjective labeled data. Due to the lack of datasets with subjective rating labels, we try to objectively measure the beauty of music. We adopt traditional measure to quantify the beauty of music.
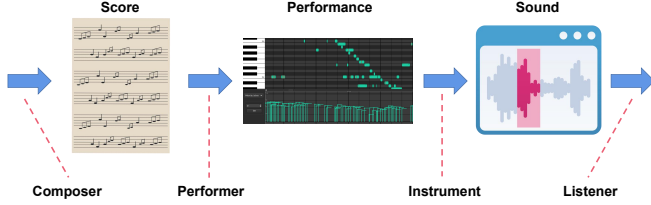
Figure 2: Three steps from creation to hearing for a music art work, and this paper focuses on the sound stage.

In this paper, Birkhoff's method [5] was selected to conduct a study of aesthetic quality assessment of music. The reason why we choose Birkhoff's aesthetic measure is that it is not only famous, but also convenient to formalize, and has a relatively stronger interpretability for beauty (perhaps the most important reason). Birkhoff formalizes the aesthetic measure of an object into the quotient between order and complexity:

$$M = \frac{O}{C} \tag{1}$$

The taste and music creation of users are highly dependent on various factors, so traditional music recommendation systems often cannot meet users' needs and cannot provide good recommendations. Meeting users' music and entertainment needs requires considering their contextual [2] and good interactive information [38]. Therefore, what we focus on is that our recommendation method should not only make recommendations more accurate but also recommend more aesthetically pleasing music to users.

The main contributions of our work are as follows:

- We creatively put forward an objective method to measure the beauty of instrumental music and singing music.
- We propose some basic music features and 4 aesthetic music features to facilitate the following computational aesthetic evaluation research tasks.
- We apply the music aesthetics model to music recommendation tasks to recommend more aesthetically pleasing music to users, thereby enhancing their aesthetic taste.

2 RELATED WORK

2.1 Music Generation

Music generation tasks is a critical research field of music information retrieval (MIR). The music generation task [30] mainly includes three stages: score generation, performance generation and audio generation, and at present, the main methods of using deep learning to generate music include CNN [50, 70], RNN [22, 29, 65], VAE [8, 57], GAN [15, 28] and Transformer [26, 61, 72].

The contribution of WaveNet [50] in the field of audio generation is revolutionary, mainly based on its ability to generate high-quality audio waveforms, which can be used to generate audio and songs for various instruments. However, models such as WaveNet [50], SampleRNN [45], and NSynth [19] are very complex, require a large number of parameters, a long training time, and a large number of examples, and lack interpretability. So other models considered, such as Dieleman et al. [14] proposed the use of ADAs to capture long-term correlations in waveforms.

The spectral characteristics of singing are also influenced by F0 motion [71]. Previously successful singing synthesizers are based on the concatenation method [7]. Although the systems are state-of-the-art in terms of sound quality and naturalness, their flexibility is limited and it is difficult to scale or significantly improve. The statistical parameter method [53] uses a small amount of training data for model adaptation. But these systems cannot match the existing smooth frequency and time.

2.2 Music Evaluation

Objective evaluation is mainly used to assess the performance of music generation models. Objective evaluation metrics are designed to standardize the results of different models, allowing for comparison of their performance. These metrics are often statistical and can be categorized into four categories: pitch-related, rhythm-related, harmony-related, and style transfer-related [30], as used in models like MuseGAN [15]. In addition, there are specialized metrics that do not directly relate to music, such as MV2H [44], which evaluates the performance of music transcription models.

Subjective evaluation is still the best way to assess aesthetics, and its methods include listening tests and visual analysis. Common listening tests include the Turing test [11, 22] and side-by-side rating [23], which provide subjective evaluations of the music's quality and degree of innovation, as well as aspects such as harmony, rhythm, structure, coherence, and overall rating [15].

There is currently limited research related to music aesthetics evaluation. SAAM [33] is a state-of-the-art music aesthetics assessment model that provides aesthetic ratings for music scores. PAAM [34] is also a state-of-the-art music aesthetic model for evaluating the beauty of instrumental performances. Audio Oracle [17] uses Information Rate (IR) as an aesthetic measure.

2.3 Music Recommendation

Combining deep learning technology with traditional recommendation methods can effectively improve the effectiveness of music recommendation in recent years.

Van et al. [64] utilized the historical listening data of users and the audio signal data of music to project users and music into a shared hidden space, enabling them to learn the hidden representations of users and songs. Jiang et al. [31] proposed an improved algorithm based on RNN to measure the similarity between different songs. Wang et al. [66] proposed a DBN based method that combines the implicit representations of users and items into a user rating matrix for model training and improving the performance of music recommendations. Sachdeva et al. [58] proposed a novel attention neural structure that utilizes the features of these tracks to better understand users' short-term preferences.
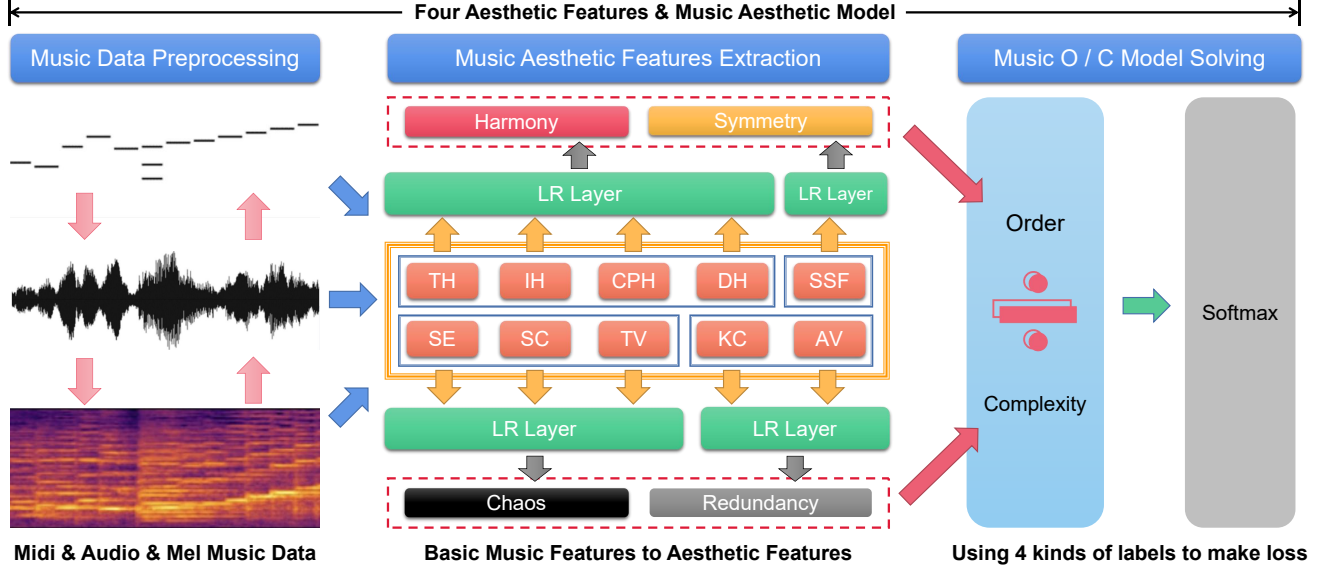
Figure 3: The overview of our aesthetic approach. We align and preprocess the midi, audio, and mel spectrogram of music, and obtain the values of ten basic music features by calculating their basic attributes. Then we put the corresponding basic music features to four logistic regression models to extract four music aesthetic features. Finally, we take the values of four music aesthetic features as the input of Birkhoff O/C model, and solve the model to get the final aesthetic score.

3 MUSIC AESTHETIC MODEL

3.1 Formalization

Based on Birkhoff's measure, we propose four aesthetic features: harmony, symmetry, chaos and redundancy. We linearly combine the order measures of molecules and the complexity measures of denominators by four logistic regression model. Fig.3 shows our method. Detailed measures explanation will be described for next. The music aesthetic measure formalization is as follows:

$$Aesthetic\ Measure = \frac{\omega_1 H + \omega_2 S + \theta_1}{\omega_3 C + \omega_4 R + \theta_2} \tag{2}$$

Where H is harmony, S is symmetry, C is chaos and R is redundancy. ω is the weight and θ is the constant.

3.2 Harmony

People feel good because there is "harmony" in music [12].

3.2.1 Timbre Harmony. The formalization of timbre harmony can be divided into two aspects: timbre balance and complementarity.

Timbre balance refers to the relative balance of sound energy in different frequency bands among instruments.

Let $S_i(f)$ denote the sound energy of the i-th instrument in frequency band f, and let F be the set of all frequency bands. Then, timbre balance can be represented as:

$$B = \max_{f \in F} \left(\frac{\max_{i=1}^n S_i(f)}{\min_{i=1}^n S_i(f)} \right) \tag{3}$$

Where n is the number of instruments, B represents the ratio of the minimum and maximum sound energy proportions among different frequency bands, and $B \in [0, 1]$.

Timbre complementarity refers to the mutual complementation of sound colors among instruments.

Let $s_i(t)$ denote the sound waveform of the i-th instrument at time t, and let $s(t)$ denote the total sound waveform of the ensemble. Then, timbre complementarity can be represented as:

$$C = \frac{1}{T} \int_0^T \sum_{i=1}^n s_i(t) s(t) dt \tag{4}$$

Where T is the total length of the music, and C represents the timbre complementarity score among instruments, with $C \in [-1, 1]$.

$$Timbre\ Harmony = \alpha \cdot B + (1 - \alpha) \cdot C \tag{5}$$

Timbre harmony is obtained by linearly weighting balance and complementarity. Where α is the weighting factor.

3.2.2 Interval Harmony. The term "interval" refers to the distance between two notes in music. An interval of 12 semitones is called an octave. Mathematical and physical studies have demonstrated that when two sound frequencies have a simple integer ratio, they produce a more pleasing auditory experience when played together. To measure the harmony of intervals, We refer to the SAAM [33] method and propose the following formula.

$$Interval\ Harmony = \sum_{i=1}^{12} \alpha_i * pir_i + \theta_{ih} \tag{6}$$

In this formula, α_i represents the weight of each interval, pir_i is the ratio of the specific interval to the total number of intervals, and θ_{ih} is a constant term.

3.2.3 Chord Progression Harmony. In tonal music, harmony progression is a specific harmonic range in chords connection.

In this study, we adopt the method in SAAM [33], which calculates the average value of progression tension to obtain a quantitative chord progress harmony. The formula is as follows:

$$Chord\ Progression\ Harmony = \lambda_1 d_1(T_i, T_{i-1}) +$$
$$\lambda_2 d_2(T_i, T_{key}) + \lambda_3 d_3(T_i - T_{key}, T_f) + \lambda_4 c(T_i) + \quad (7)$$
$$\lambda_5 m(T_i, P) + \lambda_6 h(T_i, P)$$

Here, T_i represents the i-th chord of the progression P, and λ is the weight parameter. Further details on the parameters c, m, and h can be found in [49].

3.2.4 Dynamic Harmony. According to GTTM theory [39], the metric structure of a piece of music can be derived from it. The strength of playing each note can be described as dynamic. The metrical structure of music represents the strong and weak beats.

We refer to the PAAM [34] approach to quantify the harmony of dynamics. The formula for dynamic harmony is:

$$Dynamic\ Harmony = \frac{\sum_{i=1}^n D_i \times M_i}{\sqrt{\sum_{i=1}^n (D_i)^2} \times \sqrt{\sum_{i=1}^n (M_i)^2}} \quad (8)$$

Here, i represents the i-th note at the beginning of the bar, n is the total number of aligned notes at the beginning of the bar, D represents the vector of dynamic values, and M represents the vector of metrical structure values.

3.3 Symmetry

Symmetry is considered to be the key to the perfection of art [52].

3.3.1 Self-Similarity-Fitness. In music, symmetry usually refers to a form of regularity, repetition, or mirroring in the music. In the music generation field, the structure of a piece is often regarded as a crucial aspect. Repetitive patterns can be found in almost all kinds of music. Our aim is to find how the repetitive structure affects the beauty of music.

Drawing inspiration from the aesthetics of art and images [1], we hypothesize that the beauty of music originates from the symmetry present in its compositions. Thus, we adopt Müller's fitness method [47] to gauge the extent of repetition in a musical piece. The fitness formula is expressed as follows:

$$Self\ Similarity\ Fitness = 2 \cdot \frac{\bar{\sigma}(\alpha) \cdot \bar{\gamma}(\alpha)}{\bar{\sigma}(\alpha) + \bar{\gamma}(\alpha)} \quad (9)$$

Where $\bar{\sigma}(\alpha)$ and $\bar{\gamma}(\alpha)$ refer to Müller's method [47].

3.4 Chaos

When people deal with complicated information, they will feel uncomfortable and uneasy [3].

3.4.1 Shannon Entropy. We employ Shannon entropy to quantify the degree of chaos or disorder in the internal state of a system, using histogram entropy as a specific instance of this measure. Given a finite set Ω and a random variable X taking values in Ω, with each value $x \in \Omega$ having a probability distribution $p(x) =$

$Pr[X = x]$. What we need to consider is the entropy of the music attribute histogram. The Shannon entropy $H(X)$ is defined as:

$$Shannon\ Entropy = -\sum_{x \in \Omega} p(x) \log(x) \quad (10)$$

The Shannon entropy is a widely used method to assess the average uncertainty of a random variable X.

3.4.2 Spectral Complexity. Spectral complexity is a method for measuring the complexity of a musical signal's spectral content. It is based on the spectral flux of the signal. It is a measure of how much the spectral content of a signal changes over time.

The spectral complexity of the signal is then defined as the weighted sum of the spectral flux:

$$Spectral\ Complexity = \frac{1}{T} \sum_{t=1}^{T} \sum_{f=1}^{F} |X(f,t) - X(f,t-1)| \quad (11)$$

Where $x(t)$ is a signal with N samples, $X(f,t)$ is its short-time Fourier transform, F is the number of frequency bins, T is the number of analysis frames.

3.4.3 Timbre Variability. Timbre variability is a method used to measure the diversity of timbre in music. It can calculate the degree of timbre variation in the music.

Formally, let $X = x_1, x_2, ..., x_n$ represent the n notes in the music, $f(x_i)$ represent the basic timbre feature of the ith note, and $Var(x_i)$ represent the timbre variability of the ith note. Then the Timbre variability of the entire piece of music can be represented as:

$$Timbre\ Variability = \frac{\sum_{i=1}^n Var(x_i) \cdot w_i}{\sum_{i=1}^n w_i} \quad (12)$$

Here, w_i represents the weight of the ith note.

3.5 Redundancy

Aesthetically speaking, redundancy makes people feel dull, resulting in negative emotions [40].

3.5.1 Kolmogorov Complexity. The Kolmogorov complexity $K(s)$ of a string s is defined as the length of the shortest program that outputs s on a computer. In the context of music, lossless compression can be used to estimate the redundancy of a piece of music. This can be expressed using the following formula:

$$Kolmogorov\ Complexity = 1 - \frac{K}{I_m} \quad (13)$$

Where K is the amount of information after lossless compression of music, and I_m is the original amount of information of music.

3.5.2 Autocorrelation Value. The autocorrelation value reflects the degree of redundancy in music.

Given a music signal $x(t)$, the autocorrelation value can be calculated as:

$$Autocorrelation\ Value = \frac{1}{N} \sum_{i=0}^{N-1} \frac{\sum_{t=0}^{T-i-1} x(t)x(t+i)}{\sum_{t=0}^{T-1} x^2(t)} \quad (14)$$

Where N is the maximum time lag for calculating the autocorrelation, T is the length of the music signal, and $x(t)$ is the amplitude of the signal at time t.
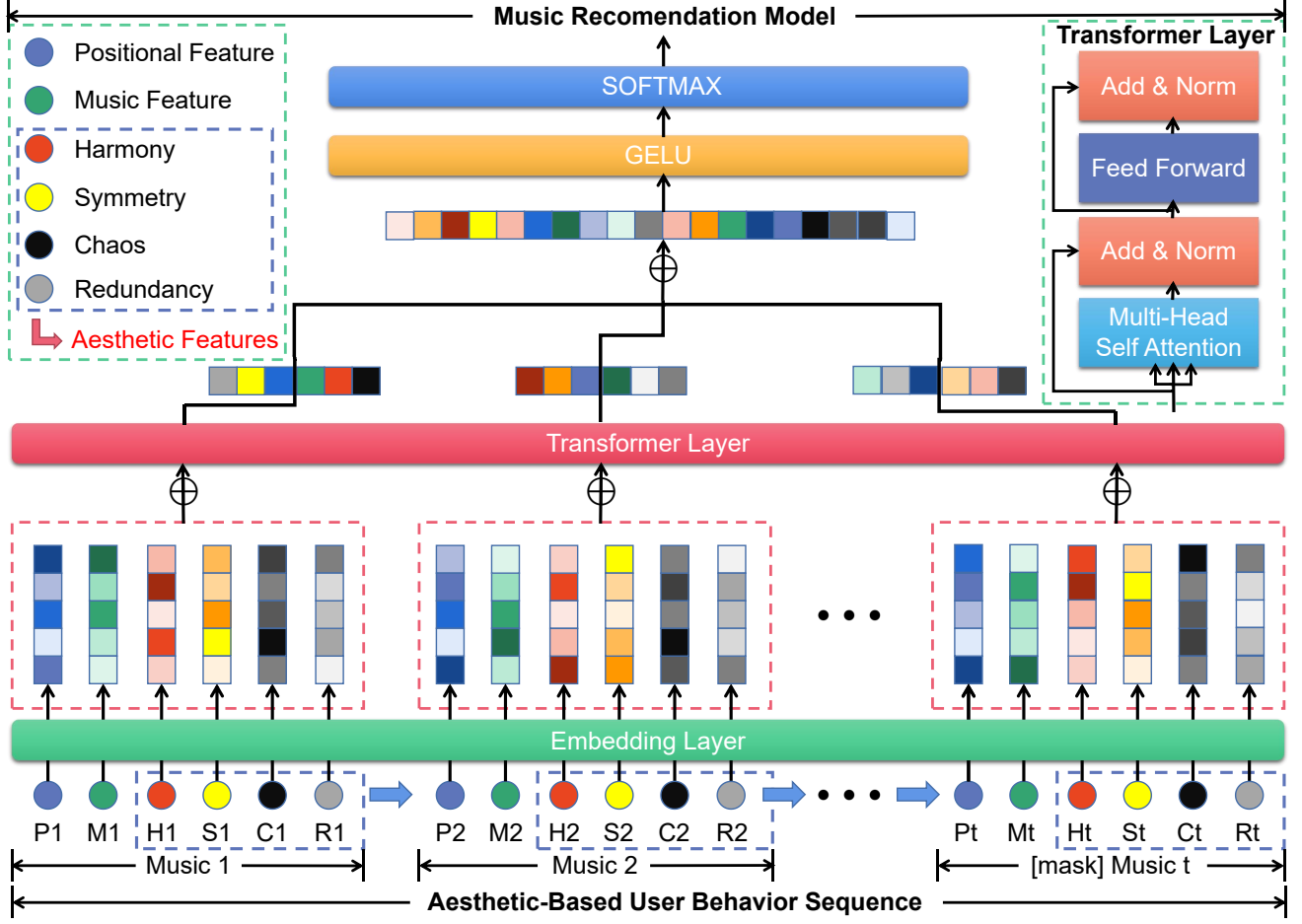
Figure 4: The overview of our recommendation approach. In addition to traditional positional features and music features, we will also extract good aesthetic features as important reference standards for sequential music recommendations. Moreover, we also consider the aesthetic rating score of music as a label value in the recommendation of sequential music.

4 AESTHETIC-AWARE RECOMMENDATION

We use CL4SRec [69] as the backbone network of our method, which is the state-of-art model for sequence recommendation tasks.

4.1 Embedding Layer

4.1.1 Positional Embedding. We transform positional features into vector representations. We assign a unique number to each position in the sequence and map each of these to a d-dimensional vector space to obtain position's vector representation.

Specifically, we use a matrix E_{pos} of size $L \times d$ to represent the vector representations of all positions, where E_{pos_i} represents the vector representation for position i. We map the position identifier to the i-th row of this matrix, E_{pos_i}.

Then, we obtain the representation for each position $\mathbf{X}_{pos}$. Then we will discuss the embedding methods for other features.

4.1.2 Other Features Embedding. For the music features and the four aesthetic features, we can use corresponding embedding layers to convert them into fixed-length vector representations.

Specifically, we use an embedding matrix E_{music} of size D_{music} to represent the vector representation of music features (CNN), where D_{music} is the embedding dimension. For each music feature j, we represent it as a vector $\mathbf{v}j \in \mathbb{R}^{Dmusic}$, and then map it to a row of the embedding matrix E_{music}, i.e., $E_{music_j} = \mathbf{v}_j$.

Similarly, we use four embedding matrices E_{aes} of size D_{aes} to represent the vector representation of aesthetic features.

Then, we obtain the representation for music feature $\mathbf{X}_{music}$ and aesthetic features $\mathbf{X}aes$.

4.1.3 Embedding Overview. We add these vector representations with the position embedding vector representations to form the input matrix $\mathbf{X}$ of the Transformer model:

$$\mathbf{X} = \mathbf{X}^{(0)} + \mathbf{X}pos + \mathbf{X}music + \mathbf{X}aes \tag{15}$$

Where $\mathbf{X}^{(0)}$ represents the vector representation for the original input sequence, $\mathbf{X}pos$ is the position embedding matrix, and $\mathbf{X}music$ and $\mathbf{X}aes$ are the embedding matrices for music and aesthetic features, respectively.

4.2 Transformer Layer

4.2.1 Multi-Head Self-Attention. For each attention head h, we compute the attention weights $\mathbf{A}^h$ as follows:

$$\mathbf{A}^h = \mathbf{softmax}\left(\frac{QW_i^h(KW_i^h)^T}{\sqrt{d_k}}\right) \tag{16}$$

Where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d_k}$ are the query, key, and value matrices obtained by linearly projecting the input sequence $\mathbf{X}$ with learned weight matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}v \in \mathbb{R}^{d \times d_k}$.

The output of the Multi-Head Self-Attention layer is then obtained by concatenating the h attention heads and passing the result through a linear layer:

$$\mathbf{MultiHead(Q, K, V)} = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)\mathbf{W}_o \tag{17}$$

$$\text{head}_i = \mathbf{Attention}(Q\mathbf{W}_i^q, K\mathbf{W}_i^k, V\mathbf{W}_i^v) \tag{18}$$

Where $\mathbf{W}o \in \mathbb{R}^{hd_v \times d}$ is a learned weight matrix.

4.2.2 Position-wise Feed-Forward Network. The feed-forward neural network consists of two fully connected layers with an activation function applied after each layer. It can be written as:

$$\mathbf{FFN(x)} = \mathbf{GELU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{19}$$

$$\mathbf{GELU(x)} = x\Phi(x) \tag{20}$$

Where $\mathbf{W}_1$ and $\mathbf{W}_2$ are both learnable weight matrices, $\mathbf{b}1$ and $\mathbf{b}2$ are the bias terms for the first and second layers, respectively.

4.2.3 Stacking Transformer Layer. The layer consists of L identical layers. Each layer applies a multi-head self-attention mechanism followed by a position-wise feed-forward network (FFN) to the input sequence. Additionally, each layer employs residual connections and layer normalization to stabilize the training process.

The output of the l-th layer is given by:

$$\begin{aligned}\mathbf{X_l} = \mathbf{LayerNorm}\big(\mathbf{X_{l-1}} + \mathbf{MultiHead}(\mathbf{LayerNorm}(\mathbf{X_{l-1}}))\big) + \\ \mathbf{FFN}\big(\mathbf{LayerNorm}(\mathbf{X_{l-1}} + \mathbf{MultiHead}(\mathbf{LayerNorm}(\mathbf{X_{l-1}})))\big)\end{aligned} \tag{21}$$

Where $l \in 1, 2, ..., L$ indicates the l-th layer of the stacking transformer layer. LayerNorm($\cdot$) represents the layer normalization operation, MultiHead($\cdot$) represents the multi-head self-attention, and FFN($\cdot$) represents the position-wise feed-forward network.

4.3 Output Layer

For the music sequence recommendation task, the output layer can be formulated as follows: given the final output H_L from the L layers of the hierarchical exchange information, we mask the item v_t at time step t and then use h_{Lt} to predict the masked item v_t. We apply a two-layer feed-forward network with GELU activation in between to produce an output distribution over target items:

$$\mathbf{P(v_t)} = \mathbf{softmax}\left(W_P \cdot \text{GELU}\left(h_{Lt} \cdot E\right) + b_P\right) + b_O \tag{22}$$

Where $\mathbf{W}_P$ is the learnable projection matrix, $\mathbf{b}_P$ and $\mathbf{b}_O$ are bias terms, and $\mathbf{E} \in \mathbb{R}^{|V| \times d}$ is the embedding matrix for the item set V.

4.4 Model Learning

We train the model using the randomly masked language modeling (LM) technique, where we randomly mask portions of song IDs and train the model to predict the original value of the masked song IDs using the context about non-masked song IDs in the sequence:

$$Input\ layer[s1,s2,s3,s4,s5] \rightarrow [s1,s2, [mask1],s4, [mask2]] \tag{23}$$

The cross-entropy function can be represented as follows:

$$J = -\frac{1}{N}\sum_{i=1}^{N} \log P(v_{ti}|v_{1i}, ..., v_{t-1,i}, v_{t+1,i}, ..., v_{ni}) \tag{24}$$

Here, N represents the number of training samples, and v_{ti} represents the masked music item in the i-th training sample.

5 EXPERIMENTS

5.1 Datasets & Setup

5.1.1 Dataset For Aesthetic Model. The dataset is divided into three types of samples: real audio played or sung by humans (positive), performance midi rendered audio (medium), and audio generated by various AI models (negative), 7:3 for training and testing.

Specifically, our positive sample uses MedleyDB [6], our mid value is audio rendered (by MuseScore3) from POP909's [67] piano and Lakh MIDI Dataset's [54] vocal performance midi, and the negative sample uses Music Transformer [27] to generate performance midi, and then use GanSynth [18] to synthesize it into audio.

AI (negative)	Rendered (medium)	Human (positive)
122	140	128

**Table 1: The number of samples in the aesthetic model.**

5.1.2 Recommendation Datasets. We select three datasets:
- MSD [4] dataset is a collection of metadata and feature analysis data for one million songs, including information about artists, albums, tracks, lyrics, genres, tonality, and more.
- Last. fm-1k [10] dataset contains implicit feedback data from users on music, including music play records, music tags, user information, and more.
- Last. fm-360k [9] dataset adds user-artist relationship and user information on top of the 1k dataset.

We clean the dataset by removing playback events from the session information that lack user or information, and only keeping session information with playback events greater than 10. Then we crawl corresponding audio source files according to playlists, and use Onsets and Frames [24] to get the corresponding MIDI.

5.1.3 Evaluation Metrics. For the aesthetic model scoring task, we evaluate the performance of the classification model using classification accuracy and F1 measure. For recommended tasks, we use evaluation metrics to assess the ranking list generated by the models, which include Hit Ratio (HR), Normalized Discounted Cumulative Gain (NDCG). As we only have one true item for each user, HR@k is equivalent to Recall@k and proportional to Precision@k. Additionally, we report HR and NDCG at k = 1, 5, and 10, where a higher value indicates better performance.
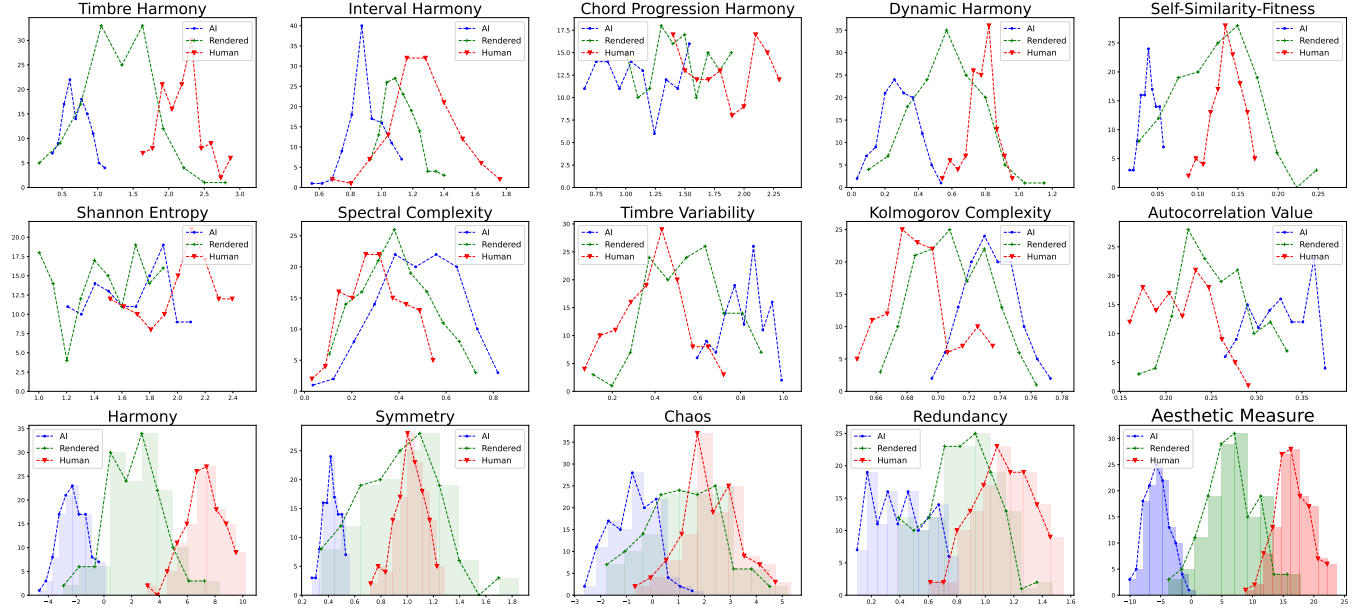
Figure 5: The first two rows in the figure show the distribution of ten basic music features on three sample sets. The last graph in the third row shows the aesthetic score distribution of our trained model on the three sample sets. The remaining four graphs in the third row show the distribution of four aesthetic features on the three sample sets.

5.2 Implementation Details

5.2.1 Music Aesthetic Model. We use music21[1] to obtain music attributes, Musescore3[2] to render music, and jSymbolic [42], jAudio [43], and librosa [41] to extract corresponding music features to complete the calculation of basic music features. For calculation, we use the aligned midi, audio, and mel. Regarding the four aesthetic features, we will linearly combine the corresponding basic music features and use four logistic regression models for pre training to obtain aesthetic features. Finally, input the value of aesthetic features into the O/C model, predict the label, select the cross entropy as the loss function, and use the Adam [36] optimizer for gradient descent. We set the learning rate to 5e-5, and after 1000 iterations, the loss function converges. For more details, please refer to the implementation of SAAM [33] and PAAM [34].

5.2.2 Music Recommendation Model. We use word2vec to process positional features, CNN to process music features, and for four aesthetic features, we use the original vector. We train the model using TensorFlow with truncated normal distribution initialization in the range of [-0.01, 0.01] for all parameters. The optimization is performed using the Adam [36] with a learning rate of 1e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, ℓ_2 weight decay of 0.01, and linear decay of the learning rate. The gradient is clipped when its ℓ_2 norm exceeded a threshold of 5. To ensure fair comparison, we set the layer number $L = 2$ and head number $h = 4$. We select the mask proportion ρ through validation, resulting in $\rho = 0.6$ for MSD, and $\rho = 0.4$ for Last.fm-1k and Last.fm-360k. Model is trained on NVIDIA GeForce GTX 1080Ti GPU with a batch size of 256.

---

[1] https://github.com/cuthbertLab/music21
[2] https://musescore.org/

5.3 Results Analysis & Discussion

5.3.1 Music Aesthetic Model. The distribution of various indicators and the final aesthetic measure in the aesthetic model are shown in Fig.5. After training our O/C model, we observe the intersection of the aesthetic scores distributions (in the last of the third row in Fig. 5) is very small, indicating the O/C model's necessity. To be more specific, TH, IH, and DH have achieved good differentiation effects, proving that human music audio is significantly superior to rendered audio and AI generated audio in terms of harmony. As for the symmetry aspect, the audio generated by AI has a significantly lower repeatability and appears to be structurally disorganized, while the differentiation between the other two is not as significant. In terms of Chaos, the differentiation of basic music features is not particularly significant. In terms of redundancy, it is clear that human audio is the best, followed by the other two. The accuracy of our model on the test set is 92.3%, and the f1 measure is 90.5%.

5.3.2 Music Recommendation Model. In Table2, we present a summary of the best performance achieved by all models on four benchmark datasets. As the NDCG@1 results are equivalent to HR@1 in our experiments, we have omitted them from the table. We have also applied some other networks (7 baseline models) as backbone networks, namely BPR [55], CDAE [68], FPMC [56], GRU4Rec [25], Caser [63], SASRec [35], Bert4Rec [62] and the performance comparison will be provided in Table2. The experimental results show that the performance is best when using CL4SRec as the backbone network, CL4SRec is superior in nearly all metrics. Note that in the results of Table2, we only add aes features on CL4SRec. The penultimate line is CL4SRec without aesthetic features, and our performance has slightly improved after integrating aesthetic features, proving the aesthetic features are meaningful.

Datasets	Metric	BPR	CDAE	FPMC	GRU4Rec	Caser	SASRec	BERT4Rec	CL4SRec	CL4SRec w/ aes
MSD	HR@1	0.0465	0.0479	0.0474	0.0493	0.0577	0.0703	0.0714	**0.0936**	<u>0.0927</u>
	HR@5	0.1182	0.1248	0.1242	0.1319	0.1681	0.1692	0.1828	<u>0.2179</u>	**0.2185**
	HR@10	0.1978	0.2014	0.2148	0.2369	0.2654	0.2687	0.2584	<u>0.3071</u>	**0.3089**
	NDCG@5	0.0941	0.0754	0.0831	0.0768	0.1146	0.1195	0.1282	<u>0.1583</u>	**0.1624**
	NDCG@10	0.1023	0.1273	0.1252	0.1342	0.1564	0.1532	0.1591	**0.1902**	<u>0.1895</u>
Last.fm-1k	HR@1	0.0443	0.0517	0.0392	0.0374	0.0633	0.0669	0.0704	<u>0.0982</u>	**0.1014**
	HR@5	0.1138	0.1532	0.1309	0.1414	0.1795	0.1863	0.2384	<u>0.2709</u>	**0.2718**
	HR@10	0.2042	0.2251	0.2389	0.2446	0.2647	0.2986	0.3613	<u>0.3999</u>	**0.4026**
	NDCG@5	0.1045	0.0985	0.1079	0.0963	0.1408	0.1675	0.1501	<u>0.1821</u>	**0.1902**
	NDCG@10	0.1152	0.1203	0.1269	0.1291	0.1832	0.1859	0.1952	**0.2254**	<u>0.2202</u>
Last.fm-360k	HR@1	0.0768	0.1248	0.1275	0.1481	0.1535	0.1858	0.2151	<u>0.2854</u>	**0.2936**
	HR@5	0.2984	0.3095	0.3730	0.4118	0.4528	0.4890	0.5272	<u>0.5894</u>	**0.5931**
	HR@10	0.4109	0.4627	0.5219	0.5809	0.5976	0.6012	0.6467	<u>0.6972</u>	**0.6981**
	NDCG@5	0.1879	0.2173	0.2418	0.3012	0.3623	0.3471	0.3767	**0.4484**	<u>0.4313</u>
	NDCG@10	0.2164	0.2579	0.2841	0.3448	0.3687	0.3512	0.4176	<u>0.4807</u>	**0.4984**

Table 2: The performance of various methods on next-item prediction is compared, with the best scores in each row shown in bold and the second-best scores underlined. Note that, we only add aes features on CL4SRec, which achieves best performance.

5.4 Subjective Evaluation

We refer to this work [16]. To confirm the effectiveness of our aesthetic model and recommendation algorithm, we conduct a subjective evaluation experiment and invite 15 professional musicians to participate. It is necessary to confirm the validity of the dataset, that is, whether the audio generated by AI has the lowest aesthetic value, followed by rendered audio, and the audio actually performed or sung by humans has the highest aesthetic value. We select 5 songs from each of the above three sample sets and ask volunteers to rate their aesthetic value. The results are shown in Table3:

AI (negative)	Rendered (medium)	Human (positive)
2.32 ± 0.27	3.25 ± 0.32	4.18 ± 0.16

Table 3: Results of listening test 1. It presents the Mean Opinion Scores (MOS) along with a 95% confidence interval.

After proving the validity of our dataset, then we select five different songs corresponding to the model rating of five levels and ask volunteers to subjectively rate them, and this will determine whether our aesthetic model can have good musical aesthetic differentiation ability, shown as Table4:

Low	Medium Low	Medium	Medium High	High
1.46 ± 0.17	2.65 ± 0.32	3.24 ± 0.27	3.82 ± 0.31	4.32 ± 0.26

Table 4: Results of listening test 2. It presents the Mean Opinion Scores (MOS) along with a 95% confidence interval.

The subjective experimental results demonstrate that our aesthetic model is effective and can measure the beauty of music. As for the recommendation algorithm, we also briefly ask volunteers to conduct rating experiments, and 93.3% of volunteers believe that our recommendation algorithm can recommend more aesthetically pleasing music, proving its effectiveness.

5.5 Ablation Study

We conduct ablation study to remove harmony, symmetry, chaos, and redundancy to train four different models. We compare them with our full model. The following Table5 shows the results:

Our Full Model	0.674
-w/o harmony	1.028
-w/o symmetry	0.723
-w/o chaos	0.785
-w/o redundancy	0.836

Table 5: Comparisons of the final MSE between the ablation models and the ground truths, in log scales.

Experiments have shown that our full model has the best performance, with harmony being the most important aesthetic feature. Regarding the performance ablation of the model, after removing aes features, there is no significant performance decline .

6 CONCLUSION

In this article, we propose an O/C based music aesthetic evaluation model and propose some music features. In addition, we also integrate our music aesthetic features for music aesthetic recommendation task. Nevertheless, there are still shortcomings in this work, such as the lack of consideration of "innovation". We hope that this article may help improve the AIGC music generation task.

ACKNOWLEDGMENTS

REFERENCES

[1] Mohammad Majid al Rifaie, Anna Ursyn, Robert Zimmer, and Mohammad Ali Javaheri Javid. 2017. On symmetry, aesthetics and quantifying symmetrical complexity. In *Computational Intelligence in Music, Sound, Art and Design: 6th International Conference, EvoMUSART 2017, Amsterdam, The Netherlands, April 19–21, 2017, Proceedings 6*. Springer, 17–32.

[2] Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Aykan Aydin, Karl-Heinz Lüke, and Roland Schwaiger. 2011. Incarmusic: Context-aware music recommendations in a car. In *E-Commerce and Web Technologies: 12th International Conference, EC-Web 2011, Toulouse, France, August 30-September 1, 2011. Proceedings 12*. Springer, 89–100.

[3] Max Bense. 1960. Programmierung des Schönen Allgemeine Texttheorie Und Textästhetik. (1960).

[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

[5] George David Birkhoff. 2013. Aesthetic measure. In *Aesthetic Measure*. Harvard University Press.

[6] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. 2014. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. *International Society for Music Information Retrieval Conference* (2014). https://www.upf.edu/web/mtg/multitrack-databases

[7] Jordi Bonada, Martí Umbert Morist, and Merlijn Blaauw. 2016. Expressive singing synthesis based on unit selection for the singing synthesis challenge 2016. *Morgan N, editor. Interspeech 2016; 2016 Sep 8-12; San Francisco, CA.[place unknown]: ISCA; 2016. p. 1230-4.* (2016).

[8] Gino Brunner, Andres Konrad, Yuyi Wang, and Roger Wattenhofer. 2018. MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer. *arXiv preprint arXiv:1809.07600* (2018).

[9] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2011. 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). *Proceedings of the 5th ACM Conference on Recommender Systems* (2011), 387–388. https://doi.org/10.1145/2043932.2044017

[10] Òscar Celma. 2008. Music Recommendation and Discovery in the Long Tail. In *Proceedings of the 1st ACM International Conference on Recommender Systems*. 5–12. https://doi.org/10.1145/1454008.1454013

[11] Ondřej Cífka, Umut Şimşekli, and Gaël Richard. 2019. Supervised symbolic music style translation using synthetic data. *arXiv preprint arXiv:1907.02265* (2019).

[12] Ana Clemente, Marcus T Pearce, and Marcos Nadal. 2022. Musical aesthetic sensitivity. *Psychology of Aesthetics, Creativity, and the Arts* 16, 1 (2022), 58.

[13] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34, 4 (2017), 80–106.

[14] Sander Dieleman, Aaron van den Oord, and Karen Simonyan. 2018. The challenge of realistic music generation: modelling raw audio at scale. *Advances in Neural Information Processing Systems* 31 (2018).

[15] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[16] Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley. 2022. Deep performer: Score-to-audio music performance synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 951–955.

[17] Shlomo Dubnov, Gérard Assayag, and Arshia Cont. 2011. Audio oracle analysis of musical information rate. In *2011 ieee fifth international conference on semantic computing*. IEEE, 567–571.

[18] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710* (2019).

[19] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*. PMLR, 1068–1077.

[20] Philip Galanter. 2012. Computational aesthetic evaluation: steps towards machine creativity. In *ACM SIGGRAPH 2012 Courses*. 1–162.

[21] Philip Galanter. 2013. Computational aesthetic evaluation: Automated fitness functions for evolutionary art, design, and music. In *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*. 1005–1038.

[22] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. 2017. Deepbach: a steerable model for bach chorales generation. In *International Conference on Machine Learning*. PMLR, 1362–1371.

[23] Albert Haque, Michelle Guo, and Prateek Verma. 2018. Conditional end-to-end audio transforms. *arXiv preprint arXiv:1804.00047* (2018).

[24] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. 2017. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153* (2017).

[25] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[26] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *arXiv preprint arXiv:1809.04281* (2018).

[27] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. [n. d.]. Music Transformer: Generating Music with Long-Term Structure. In *International Conference on Learning Representations*.

[28] Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, and Zhefeng Wang. 2022. Singgan: Generative adversarial network for high-fidelity singing voice generation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2525–2535.

[29] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam. 2019. VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance.. In *ISMIR*. 908–915.

[30] Shulei Ji, Jing Luo, and Xinyu Yang. 2020. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801* (2020).

[31] Miao Jiang, Ziyi Yang, and Chen Zhao. 2017. What to play next? A RNN-based music recommendation system. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE, 356–358.

[32] Xin Jin, Le Wu, Geng Zhao, Xiaodong Li, Xiaokun Zhang, Shiming Ge, Dongqing Zou, Bin Zhou, and Xinghui Zhou. 2019. Aesthetic attributes assessment of images. In *Proceedings of the 27th ACM international conference on multimedia*. 311–319.

[33] Xin Jin, Wu Zhou, Jinyu Wang, Duo Xu, Yiqing Rong, and Shuai Cui. 2023. An Order-Complexity Model for Aesthetic Quality Assessment of Symbolic Homophony Music Scores. *arXiv preprint arXiv:2301.05908* (2023).

[34] Xin Jin, Wu Zhou, Jinyu Wang, Duo Xu, Yiqing Rong, and Jialin Sun. 2023. An Order-Complexity Model for Aesthetic Quality Assessment of Homophony Music Performance. *arXiv preprint arXiv:2304.11521* (2023).

[35] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[36] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[37] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 662–679.

[38] Saurabh Kulkarni and Sunil F Rodd. 2020. Context Aware Recommendation Systems: A review of the state of the art techniques. *Computer Science Review* 37 (2020), 100255.

[39] Fred Lerdahl and Ray S Jackendoff. 1996. *A Generative Theory of Tonal Music, reissue, with a new preface*. MIT press.

[40] Ruth Lorand. 2002. *Aesthetic order: a philosophy of order, beauty and art*. Routledge.

[41] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. 18–25.

[42] Cory McKay, Julie Cumming, and Ichiro Fujinaga. 2018. JSYMBOLIC 2.2: Extracting Features from Symbolic Music for use in Musicological and MIR Research.. In *ISMIR*. 348–354.

[43] Cory McKay, Ichiro Fujinaga, and Philippe Depalle. 2005. jAudio: A feature extraction library. In *Proceedings of the international conference on music information retrieval*, Vol. 11. 600–3.

[44] Andrew McLeod and Mark Steedman. 2018. Evaluating Automatic Polyphonic Music Transcription.. In *ISMIR*. 42–49.

[45] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2016. SampleRNN: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837* (2016).

[46] Leonard B Meyer. 1956. Meaning and emotion in music.

[47] Meinard Müller, Nanzhu Jiang, and Peter Grosche. 2013. A Robust Fitness Measure for Capturing Repetitions in Music Recordings With Applications to Audio Thumbnailing. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 3 (2013), 531–543.

[48] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2408–2415.

[49] María Navarro-Cáceres, Marcelo Caetano, Gilberto Bernardes, Mercedes Sánchez-Barba, and Javier Merchán Sánchez-Jara. 2020. A computational model of tonal tension profile of chord progressions in the tonal interval space. *Entropy* 22, 11 (2020), 1291.

[50] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu.

2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[51] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. 2020. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications* 32 (2020), 955–967.

[52] Harold Osborne. 1986. Symmetry as an aesthetic factor. *Computers & Mathematics with Applications* 12, 1-2 (1986), 77–82.

[53] Keiichiro Oura, Ayami Mase, Tomohiko Yamada, Satoru Muto, Yoshihiko Nankaku, and Keiichi Tokuda. 2010. Recent development of the HMM-based singing voice synthesis system—Sinsy. In *Seventh ISCA Workshop on Speech Synthesis*.

[54] Colin Raffel, Ron Weiss, Johan Pauwels, William Colombo, Curtis Hawthorne, and Douglas Eck. 2016. Learning-based methods for comparing sequences, with applications to audio-to-MIDI alignment and matching. In *International Conference on Machine Learning*. PMLR, 1368–1377.

[55] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[56] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[57] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2018. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*. PMLR, 4364–4373.

[58] Noveen Sachdeva, Kartik Gupta, and Vikram Pudi. 2018. Attentive neural architecture incorporating song features for music recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 417–421.

[59] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. 2015. Music recommender systems. *Recommender systems handbook* (2015), 453–492.

[60] Tiancheng Shen, Jia Jia, Yan Li, Yihui Ma, Yaohua Bu, Hanjie Wang, Bo Chen, Tat-Seng Chua, and Wendy Hall. 2020. Peia: Personality and emotion integrated attentive model for music recommendation on social media platforms. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 206–213.

[61] Yi-Jen Shih, Shih-Lun Wu, Frank Zalkow, Meinard Muller, and Yi-Hsuan Yang. 2022. Theme transformer: Symbolic music generation with theme-conditioned transformer. *IEEE Transactions on Multimedia* (2022).

[62] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[63] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.

[64] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. *Advances in neural information processing systems* 26 (2013).

[65] Elliot Waite, Douglas Eck, Adam Roberts, and Dan Abolafia. 2016. Project Magenta: Generating long-term structure in songs and stories. *Online] https://magenta. tensorflow. org/2016/07/15/lookback-rnn-attention-rnn* (2016).

[66] Xinxi Wang and Ye Wang. 2014. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. 627–636.

[67] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Xianbin Gu, and Gus Xia. 2020. Pop909: A pop-song dataset for music arrangement generation. *arXiv preprint arXiv:2008.07142* (2020).

[68] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the ninth ACM international conference on web search and data mining*. 153–162.

[69] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.

[70] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847* (2017).

[71] Yuan-Hao Yi, Yang Ai, Zhen-Hua Ling, and Li-Rong Dai. 2019. Singing voice synthesis using deep autoregressive neural networks for acoustic modeling. *arXiv preprint arXiv:1906.08977* (2019).

[72] Xueyao Zhang, Jinchao Zhang, Yao Qiu, Li Wang, and Jie Zhou. 2022. Structure-enhanced pop music generation via harmony-aware learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1204–1213.