AN ORDER-COMPLEXITY MODEL FOR AESTHETIC QUALITY ASSESSMENT OF HOMOPHONY MUSIC PERFORMANCE

Xin Jin[1], *Wu Zhou*[1], *Jinyu Wang*[1], *Duo Xu*[2*], *Yiqing Rong*[1] *and Jialin Sun*[1]

[1]Department of Cyberspace Security, Beijing Electronic Science and Technology Institute, China
[2]Department of Art Management, Tianjin Conservatory of Music, China, many33@126.com

ABSTRACT

Althoughth computational aesthetics evaluation has made certain achievements in many fields, its research of music performance remains to be explored. At present, subjective evaluation is still a ultimate method of music aesthetics research, but it will consume a lot of human and material resources. In addition, the music performance generated by AI is still mechanical, monotonous and lacking in beauty. In order to guide the generation task of AI music performance, and to improve the performance effect of human performers, this paper uses Birkhoff's aesthetic measure to propose a method of objective measurement of beauty. The main contributions of this paper are as follows: Firstly, we put forward an objective aesthetic evaluation method to measure the music performance aesthetic; Secondly, we propose 10 basic music features and 4 aesthetic music features. Experiments show that our method performs well on performance assessment.

Index Terms— Computational aesthetics, Music performance evaluation, Birkhoff's measure, Music features

1. INTRODUCTION

Computational aesthetics evaluation [1] enables computers to make quantitative aesthetic judgments on work of arts, which usually include architecture, painting and music. It can be used to compare the aesthetic feeling of different works of art for aesthetic quality assessment [2].

Music aesthetics has always been a difficult problem, which is proposed and attempt to be solved by mathematicians (going back to Pythagoras) up to music as emotional expression [3] or music as language [4]. This study believes that it is necessary to establish an objective method to quantify aesthetic feeling of music, and the aesthetic feeling of music is not completely subjective.

Nowadays, compared with human performers, performance generated by AI sounds monotonous and lacks spirituality. For now, the machine may only have learned the distribution of a music performance attributes like dynamic and tempo, which may be one of the reasons why the performance of the machine is very mechanical and clumsy.

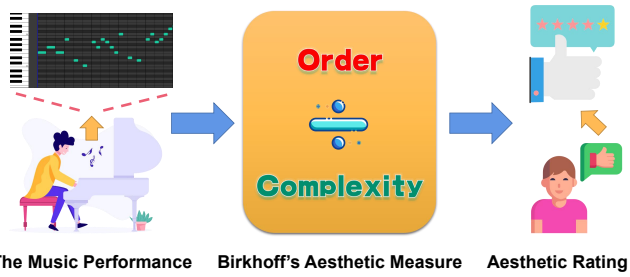

Fig. 1: The quality of music performance can be evaluated through the Performance Aesthetic Assessment Model.

Current music performance data is not labeled with aesthetic ratings like AVA [5] in the field of image aesthetic evaluation. Due to the lack of datasets with subjective rating labels, we adopt objective traditional aesthetic measure.

In this paper, Birkhoff's method [6] was selected to conduct a study of aesthetic quality assessment. The reason why we choose the method is that it has a strong interpretability for beauty. Birkhoff formalizes the aesthetic measure of an object into the quotient between order and complexity:

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

Fig1 shows the main tasks of this article. The main contributions of our work are as follows:

- We put forward an aesthetic evaluation method to measure the music performance aesthetics.
- We propose 10 basic music performance features and 4 aesthetic music features to facilitate the following music information retrieval research tasks.

2. RELATED WORK

2.1. Music Performance Generation

In this study, we focus on the tasks related to music performance generation. Maezawa et al.uses LSTM-based VAE [7] to generate expressive performance by giving conditional score, which renders performance with interpretation variations. VirtuosoNet [8] is a relatively mature work. It has

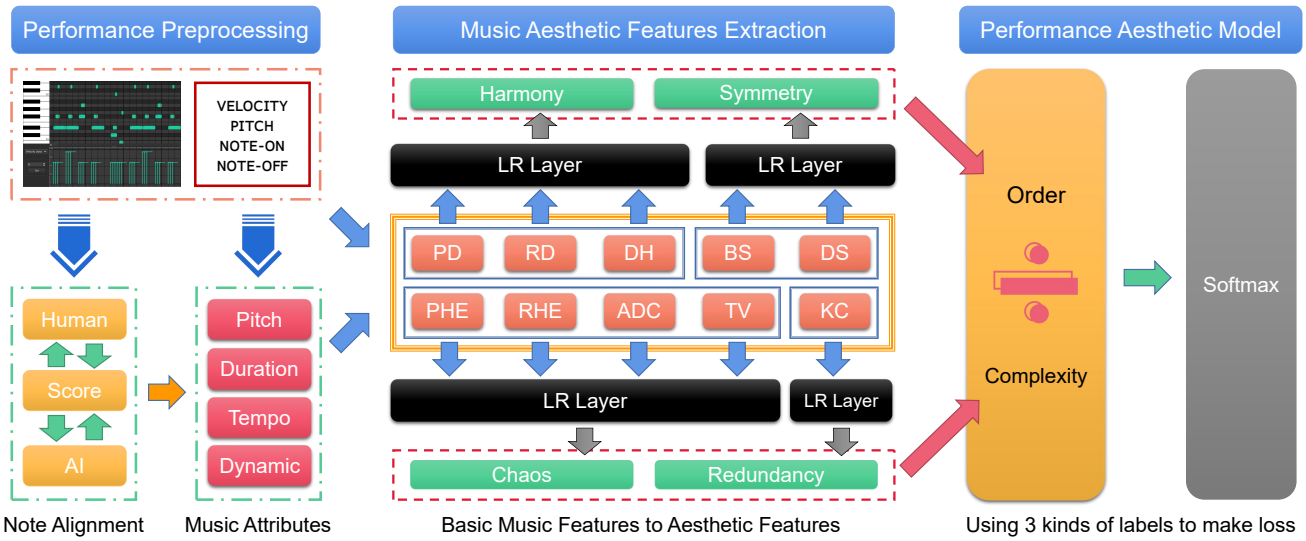

Fig. 2: The overview of our approach. First of all, we align the notes of the score and performance midi. Then, we obtain music attributes from each note after the original performance and alignment. Next, we use the formula in section 3 to process music attributes and calculate ten basic music features. After that, we put 10 basic music features into the corresponding four logical regression layers (4 black blocks in the figure), and get 4 aesthetic features. Finally, we put four aesthetic features into our aesthetic assessment model, and then establish a loss through softmax to determine the weight.

considered the important role of tempo, dynamic, and articulations in performance, which is considered to be a key issue. There are also some works such as DeepJ [9] and MIDI-VAE [10]. They also consider dynamic, but they still focus on the task of music score generation, so we will not elaborate here.

2.2. Music Evaluation

The objective evaluation metrics are designed to unify the results of different models, so that the performance of different models can be compared. These metrics are usually statistics, and they mainly include four categories: pitch-related, rhythm-related, harmony-related and style transfer-related. These metric categories are used by models like MuseGAN [11] and Jazz Transformer [12].

Subjective evaluation methods mainly include listening test and visual analysis. The main methods of listening test include turning test [13], side by side rating [14], etc. These methods make a subjective evaluation of whether music is created by machines or people, as well as the aspects of music such as the harmony competition, the rhythm, the structure, the coherence and the overall rating.

At present, there is a little work related to the evaluation of music aesthetics. Audio Oracle [15] (AO) uses Information Rate (IR) as an aesthetic measure. However, it cannot clarify what kind of specific aesthetic intention the system has. There is also a method to measure the beauty of music with Zipf's law [16], which applies the rule of word frequency to music.

3. OUR AESTHETIC ASSESSMENT 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. Detailed measures explaination will be described in sections 3.2, 3.3, 3.4 and 3.5. Fig2 shows the process of our work. The music aesthetic measure formula 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

3.2.1. Pitch Deviation & Rhythm Deviation

Pitch deviation and rhythm deviation are used to judge the quality of the pitch and rhythm accuracy played by the performer. These two features compare the pitch attribute and rhythm attribute of each note in performance with the pitch attribute and rhythm attribute of each note in the original score, so as to calculate the pitch and rhythm deviation between the performer's performance and the original score. Performers must play the notes on the music correctly. The calculation formula of deviation is given below:

$$Deviation = \frac{\sum_{i=1}^{n} |X_i - T_i|}{\sum_{i=1}^{n} |T_i|} \quad (3)$$

Where i indicates the ith note in music, n represents the total number of aligned notes, X_i represents the attribute value of the ith note of perfomance and T_i represents the attribute value of the ith note of score.

3.2.2. Dynamic Harmony

According to GTTM theory [4], given a piece of music, the metric structure of the music can be obtained. The metrical structure is the rhythmic structure in a piece of music, which well describes the strong and weak structure in music.

We use cosine similarity to measure the matching degree between dynamic and metrical structure value. The formula of dynamic harmony is as follows:

$$Dynamic\ Harmony = \frac{\sum_{i=1}^{n} D_i \times M_i}{\sqrt{\sum_{i=1}^{n} (D_i)^2 \times \sum_{i=1}^{n} (M_i)^2}} \quad (4)$$

Where i represents the ith note in the beginning of bar, n represents the total number of aligned notes which in the beginning of bar, D represents the vector of dynamic values and M represents the vector of metrical structure values.

3.3. Symmetry

3.3.1. Beat Skewness & Dynamic Skewness

Skewness is used to measure the degree of asymmetry of distribution and it is the third-order normalized moment of the sample. In performance, the two more important factors are dynamic and beat. We calculate their skewness in symmetry and take the absolute value. The formula is given below:

$$Skewness = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] \quad (5)$$

Where E represents mathematical expectation, μ represents the mean value, σ representative standard deviation.

3.4. Chaos

3.4.1. Pitch & Rhythm Histogram Entropy

We use Shannon entropy and the definition of histogram entropy is given below:

$$Histogram\ Entropy = -\sum_{x \in \Omega} p(x) \log(x) \quad (6)$$

It usually used as the measurement of the degree of chaos and uncertainty in the internal state of a system.

3.4.2. Average Dynamic Changes

There are strong (f) and weak (p) symbols in the music score to express the dynamic when playing. The calculation of average dynamic changes is as follows:

$$Average\ Dynamic\ Changes = \frac{\sum_{i=2}^{n} |D_i - D_{i-1}|}{n-1} \quad (7)$$

Where i indicates the ith note in music, n represents the total number of aligned notes. Starting from the second note, let each note make the dynamic difference between each note D_i and the dynamic of the previous note D_{i-1}, add them up and divide by n-1.

3.4.3. Tempo Variability

We design tempo variability to measure the chaos of tempo. Tempo variability is described as follows:

$$Tempo\ Variability = \sqrt{\frac{\sum_{i=1}^{n} (t_i - \bar{t})^2}{n-1}} \quad (8)$$

Tempo variability measures standard deviation of the tempo in beats per minute. Where t_i represents the ith tempo sample value $\bar{t}$ represents the mean of tempo.

3.5. Redundancy

3.5.1. Kolmogrov Complexity

For a string s, Kolmogorov complexity $K(s)$ of the string s refers to the shortest program to calculate the string s on a computer. In essence, the Kolmogorov complexity of a string is the length of the final compressed version of the string.

Lossless compression of music is a general measure of music's redundancy. We use lossless compression of music to describe it. It can be formalized as the following formula:

$$Kolmogorov\ Complexity = 1 - \frac{K}{I_m} \quad (9)$$

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

4. IMPLEMENTATION

4.1. Datasets

There are a few aligned datasets of music score and performance, we finally choose ASAP [17] as our dataset.

As shown in Table1, we plan to select the score played by the machine in ASAP as the negative sample (not beautiful performance), and then select the performance of the corresponding score in ASAP as the positive sample (beautiful performance). Then, in order to obtain an intermediate value,

Score	AI (VirtuosoNet)	Human
234	234	1060

Table 1: The number of samples for each positive, negative and intermediate value.

we use VirtuosoNet's pre-trained model [8], take ASAP score as input, and render a corresponding number of performance as the intermediate value (just so so performance).

In this way, we obtain three types of samples: the performance directly rendered by music scores, the AI generated the performance, and human performance.

4.2. Note Alignment

We need to use the attributes of each note in the score as reference values to ensure that the "improvisation" of performance is also based on evidence. To calculate features such as deviation, you need to align the notes of score and performance.

We choose a score to performance algorithm proposed by Nakamura et al. [18]. The algorithm processes notes executed asynchronously, as well as missing and additional notes in performance, and returns a list of note-to-note matches.

Although the note alignment algorithm is used, it shows that our alignment still has errors. So we adopt the refinement algorithm in VirtuosoNet [8] to improve the note alignment.

4.3. Basic Features Calculation

To calculate the basic music features, We must get the attributes of the note. We use music21[1] and jSymbolic [19] to get the pitch, duration, tempo and dynamic of the notes.

We use absolute pitch for subsequent calculation. According to the formula in 3.2.1, we calculate the deviation on the aligned notes to get PD. Then we calculate the absolute pitch histogram and use the entropy formula to calculate the PHE. Since then, the pitch related features have been calculated.

For RD, we can directly use music21 to obtain the duration attribute of the note and calculate the rhythm deviation directly. For the calculation method of RHE, we use a coding form of rhythm in jSymbolic, which encodes rhythm into 12 different lengths to eliminate the difference between performance and score to obtain a normalized rhythm histogram.

We divide velocity (dynamic) into five equal parts, setting 0-24 as level 0, 25-50 as level 1, 51-76 as level 2, 76-101 as level 3, and 102-127 as level 4. Since the reference dynamic value is labeled by bar, we take the mean of the dynamic values of all notes in each bar to form an n-dimensional vector, where n is the number of bars. The referenced dynamic is also a n-dimensional vector. Then we calculate the cosine similarity between the referenced dynamic value and the dynamic in performance to obtain the dynamic harmony (DH). Similarly, we can calculate the values of DS and ADC according to the formula in previous. Fig3 briefly shows our approach.

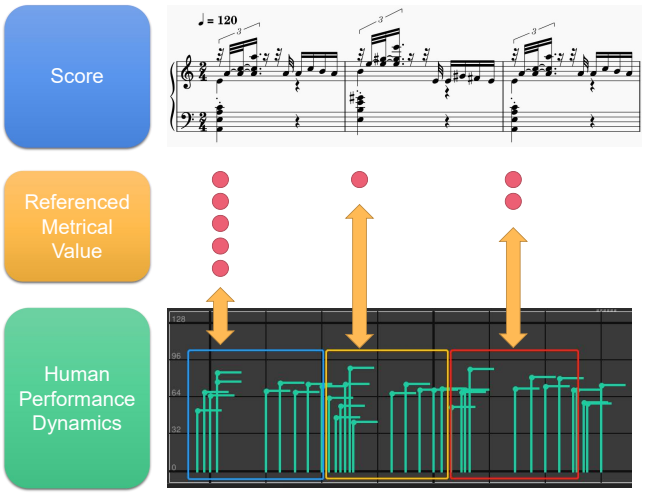

Fig. 3: Calculating the mean value of dynamic in each bar to match the referenced metrical value.

For tempo, we use the concept of beat in jSymbolic to describe the tempo of music, which is essentially another way to describe the rhythm. We use the beat histogram to calculate BS. For the feature of tempo variation, we directly choose the implementation of jSymbolic to get the value of TV.

Finally, only KC has not been calculated. We use musescore3 to render the midi files, render them into wav audio files, and sample at the frequency of 44100Hz to ensure that the music information is lossless. Then, we refer to Monkey's Audio's[2] lossless compression method to compress music into ape format, so as to calculate KC.

4.4. The Aesthetic Features & Loss Optimization

We combine our basic music features in a logistic regression model to get four aesthetic features. We normalize the value of basic music features, and then put all samples into the logistic regression model for training.

Then we put the original value of the aesthetic features into the formula in 3.1, and then calculate its softmax. We choose the cross entropy loss function, and use the gradient descent method to minimize the loss function. By setting the learning rate to 0.01, after 1000 iterations, the loss function converges and the model parameters are solved.

5. EXPERIMENTS

5.1. Main Results & Ablation Study

The main results are divided into two parts. One is to introduce feature distribution, and the other is to use the metric in section 5.1 to evaluate the performance of the model.

[1]https://web.mit.edu/music21/
[2]https://www.monkeysaudio.com/

Dataset	PD↓	RD↓	DH↑	BS↓	DS↓	PHE↓	RHE↓	ADC↑	TV↑	KC↓	H↑	S↑	C↑	R↑
Score	**0.09**	0.32	0.19	0.76	0.65	3.30	1.62	0.87	3.30	0.77	-0.77	2.71	9.40	-0.48
AI	0.12	**0.22**	0.56	0.59	0.41	**3.29**	1.44	5.25	3.20	0.75	0.56	0.91	5.15	-0.46
Human	0.13	0.29	**0.78**	**0.33**	**0.20**	3.63	**1.01**	**10.4**	**7.26**	**0.72**	**2.76**	**3.27**	**11.65**	**1.01**

Table 2: The above is the experimental data on three datasets, which are displayed by the mean value of ten basic music features and four aesthetic features. **The bold data in the table has higher aesthetic significance.** It is obvious from the table that human's performance has better aesthetic significance than the music score and the performance generated by AI.

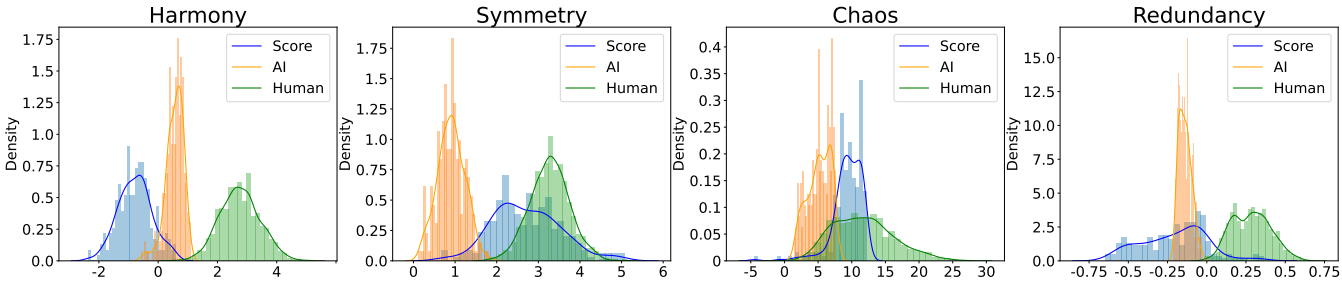

Fig. 4: The distribution of 4 aesthetic features, which are obtained through logistic regression. Among them, Harmony is better differentiated, while Symmetry, Chaos, Redundancy are worse differentiated, showing the necessity of using O/C model.

In Table2, we have calculated the mean value of 10 basic music features and 4 aesthetic features on three data sets, which makes "beauty" interpretable. Let's analyze each basic feature in detail:

Score has the lowest PD, while human's PD is the highest. According to common sense, human plays "not very accurately", but this just shows that score's playing is very mechanical, and human's playing can be more flexible when it has certain playing skills. So is RD. DH is the focus of our research. The high DH value of human shows the importance of dynamic performance to conform to the metrical structure. The value of BS and DS in human playing is significantly lower, which indicates that the beat and dynamic played by human are more symmetrical and aesthetic than AI and score. Humans' performance higher PHE and lower RHE, indicating that people may prefer to add some of their own playing skills, such as decorative tones. High ADC and TV values also indicate the diversity of human playing, and low KC values also indicate the low redundancy of human playing.

Fig4 shows the distribution of four aesthetic features. Unlike the direct use of deep learning and the use of neural networks to classify, the values generated in each step can be consulted to explain where "beauty" comes from. Fig5 shows the distribution of aesthetic scores on the three datasets. As shown in the figure, the intersection of the distribution is very small, which to proves that our model performances very well. Rating examples are shown in supplementary materials.

We also conducts ablation experiments to remove harmony, symmetry, chaos and redundancy respectively to train four different models. We compare them with the full model. We choose accuracy and kappa as the evaluation metrics. The results are shown in Table3.

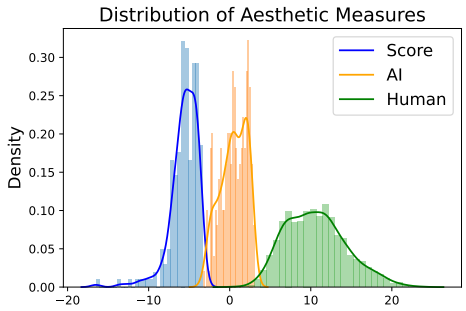

Fig. 5: The 3 distribution of Birkhoff's aesthetic measure.

5.2. Subjective Evaluation

Subjective evaluation is still the standard to test the effectiveness of aesthetic scoring. We randomly selected 15 (5 * 3) pieces of music from score, AI and human datasets, and each piece lasts about 15 seconds. Volunteers participating in the subjective experiment need to choose one in each group (5 groups) that they think is the best.

We find a total of 31 volunteers with 155 samples, and each volunteer listens to all the pieces. Among them, 32 samples (20.6%) prefer score performance, 39 samples (25.2%) prefer AI performance, and 84 samples (54.2%) prefer human performance. For detail, in the supplementary materials.

metric	w/o harmony	w/o symmetry	w/o chaos	w/o redundancy	our full model
accuracy	77.6%	88.7%	82.4%	86.5%	**92.3%**
kappa	0.579	0.791	0.653	0.726	**0.874**

Table 3: The value of kappa coefficient ranges from -1 to 1. The larger the kappa coefficient, the better the performance of the model. The results show that our full model has the best performance.

6. CONCLUSION

In summary, we propose an aesthetic model based on Birkhoff's order-complexity measure to assess the beauty of homophony music performance. For the AI music generation task, we propose 10 basic music features and 4 aesthetic features to evaluate the beauty of the generated music, which will help improve the quality of the AI music generation task. We have also done experiments on three different types of data, score, AI, and human. However, we have to admit that our model also has defects, such as the concept of symmetry, and our model does not consider "creativity", which is very important. Moreover, the aesthetic study of music audio quality assessment is worth exploring in future.

7. REFERENCES

[1] Philip Galanter, "Computational aesthetic evaluation: steps towards machine creativity," in *ACM SIGGRAPH 2012 Courses*, pp. 1–162. 2012.

[2] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14143–14152.

[3] Peter Kivy, *Sound sentiment: An essay on the musical emotions, including the complete text of the corded shell*, Temple University Press, 1989.

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

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

[6] George David Birkhoff, "Aesthetic measure," in *Aesthetic Measure*. Harvard University Press, 2013.

[7] Akira Maezawa, "Deep piano performance rendering with conditional vae," in *19th International Society for Music Information Retrieval Conference (ISMIR) Late Breaking and Demo Papers*, 2018.

[8] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, and Juhan Nam, "Virtuosonet: A hierarchical rnn-based system for modeling expressive piano performance.," in *ISMIR*, 2019, pp. 908–915.

[9] Huanru Henry Mao, Taylor Shin, and Garrison Cottrell, "Deepj: Style-specific music generation," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. IEEE, 2018, pp. 377–382.

[10] Gino Brunner, Andres Konrad, Yuyi Wang, and Roger Wattenhofer, "Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer," *arXiv preprint arXiv:1809.07600*, 2018.

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

[12] Shih-Lun Wu and Yi-Hsuan Yang, "The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures," *arXiv preprint arXiv:2008.01307*, 2020.

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

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

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

[16] Bill Manaris, Charles McCormick, and Tarsem Purewal, "Can beautiful music be recognized by computers," Tech. Rep., Technical Report CoC/CS, 2002.

[17] Francesco Foscarin, Andrew Mcleod, Philippe Rigaux, Florent Jacquemard, and Masahiko Sakai, "Asap: a dataset of aligned scores and performances for piano transcription," in *International Society for Music Information Retrieval Conference*, 2020, number CONF, pp. 534–541.

[18] Eita Nakamura, Kazuyoshi Yoshii, and Haruhiro Katayose, "Performance error detection and post-processing for fast and accurate symbolic music alignment.," in *ISMIR*, 2017, pp. 347–353.

[19] Cory McKay, Julie Cumming, and Ichiro Fujinaga, "Jsymbolic 2.2: Extracting features from symbolic music for use in musicological and mir research.," in *ISMIR*, 2018, pp. 348–354.