A quantitative aesthetic analysis of artificial intelligence generated music

Ivan Robert-Constantin
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
ivanrobertconstantin@gmail.com

Ștefan Trăușan-Matu
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
and
Research Institute for Artificial Intelligence
and
Academy of Romanian Scientists
stefan.trausan@upb.ro

ABSTRACT

In the realm of the arts, the assessment of computational aesthetics has made significant progress, but further studies in this area are still needed for music. Despite the large number of automatically generated songs, the quality of musical scores produced by artificial intelligence (AI) technology is inferior to that of music composed by human composers. Artificial intelligence often produces repetitive and emotionless musical soundtracks. Even if in the musical field the aesthetics and quality of a song may seem like a completely subjective thing, this paper presents a quantitative and objective analysis model on a set of songs both composed by musicians and intelligent agents, which can form a solid learning base for a score generation algorithm. This paper describes a suite of musical features extracted from the dataset, each with an important role in evaluating the quality of a piece.

Author keywords: aesthetics, computer-generated music, generative artificial intelligence.

DOI: 10.37789/rochi.2023.1.1.10

INTRODUCTION

Music is the art of combining vocal or instrumental sounds for beauty of form or emotional expression, usually according to cultural standards of rhythm, melody, and, in most Western music, harmony. Both the simple folk songs and the complex electronic compositions belong to the same activity, music. Both are conceptual and auditory [1], and these factors have been present in music of all styles and periods of history, all over the world [2].

For many centuries, musical creation was a human concept. In order for a new song to be created and sent through various channels to the public, a composer was needed who had enough experience and talent to know how to put the musical notes in a specific order and choose a suitable rhythm so that the listener could appreciate, resonate with the melody and find yourself in it.

Between the 1970s and 1990s, the development of artificial intelligence went through two stages of excitement and stagnation, the latter known as "AI Winters" due to the lack of computational power of that period [3]. But with the recent major technological advances, there has been a substantial development in the fields of artificial intelligence and machine learning, known as the "AI Boom" [4].

Even though there are many methods to automatically generate songs for various instruments or even full scores for an orchestra, their quality is far inferior to the quality of music written by human composers. This is probably due to the way generative machine learning models function. The intelligent agents searches its entire knowledge base for which note or set of notes is most likely to be played based on what it has generated so far or based on the user's preferences, such as musical genre, tempo, or key. Strictly statistical approaches leads to the creation of emotionless, repetitive or even disturbing melodies for listeners.

One of the simplest methods to measure the quality of a song is Birkhoff's formula for computing aesthetics in general, which is directly proportional with order and smaller complexity:

$$Aesthetic\ Measure = \frac{Order}{Complexity}$$

This method of assessing aesthetics can be used on a wide range of creative pieces from pictures to GUIs because it can be adapted to consider any relevant features such as symmetry and rhythm [5].

For music, the starting point is the premise that a song is pleasing to the ear if it has a high degree of order, i.e., the elements are symmetrical and in harmony, and if it is not too complex, i.e., there are no sudden changes in the melodic line

and there are not too many sounds played at the same time [6].

It can be seen from this scoring formula why this is not the best way to measure the quality of a song created using machine learning techniques because it would make the algorithm think that the best songs composed by it are the simple and monotonous ones, where everything is symmetrical and harmonic, and the complexity is low.

That is why this paper proposes an improvement and comes up with a deeper understanding and analysis of the aesthetics of songs in order to be able to open new doors in the field of automatic generation of songs, having a way to compose a labeled data set, suitable for supervised learning.

STATE OF THE ART

In order to be able to make a complete, fair and objective aesthetic comparison between human-generated music and AI-generated music, the same set of aesthetic metrics was applied to datasets of this both cases.

Currently, there are several datasets for computer music research, but out of the available sets, POP909 [7] and MAESTRO [8] were chosen. The decision was based on the necessity of having input data in the form of musical scores (i.e., MIDI or MusicXML formats) due to the application of techniques closely aligned with music theory concepts such as interval harmonies and pitch and rhythm distributions. This approach would not be possible, the input being in audio waves formats or compressed audio files, i.e., WAV and mp3. Besides these technical choices, these datasets provide a broader field of musical genres which helped understanding and interpreting the results, while not being constrained to a single musical genre.

The MAESTRO dataset [8], acronym for "MIDI and Audio Edited for Synchronous TRacks and Organization" represents a comprehensive collection of virtuosic piano performances, amounting to approximately 200 hours of recorded content. The dataset boasts meticulous precision, with note labels and audio waveforms aligned to a remarkable accuracy of around 3 milliseconds. This synchronization facilitates the seamless coordination between the recorded musical events and their corresponding audio representations, enabling researchers and practitioners to delve into detailed analyses and investigations of piano performance techniques, musical expression, and related areas of study.

Meanwhile, the POP909 dataset [7] contains 909 popular songs, each with multiple versions of pianos created by professional musicians. The tracks are in MIDI format, aligned to the main song (also in MIDI format) and the original audios. In addition, each song is manually tagged with metadata such as tempo, key, and chords extracted using music information retrieval algorithms. This additional data can be found attached in text files attached to each song.

This combination of genres ensures a more comprehensive understanding of how these measures can classify a song as great, irrespective of its style.

In this paper, MuseNet [9], an advanced OpenAI-developed music generation model based on artificial intelligence, was used to generate a corpus consisting of 144 original musical compositions. The corpus contains songs based on different prompts ranging from classical music in Mozart's style, to songs inspired from modern pop artists like Lady Gaga.

IMPLEMENTATION

In order to apply Birkhoff's formula, a method for quantifying both order and complexity had to be established. As a solution, a set of four features was selected for extraction from each song:

- Interval Harmony
- Symmetry
- Interval Consonance
- Shannon Entropy

Following the computation of these features, values were stored in a database. Subsequently, outliers falling below the 5th percentile and above the 95th percentile were removed, and all the data was normalized, resulting in the following formula for the ultimate assessment of aesthetic scores:

$$Aesthetic\ Measure\ = \frac{IH\ +\ S\ +IC}{3\ *\ SE}$$

Where IH stands for Interval Harmony, S stands for Symmetry, IC stands for Interval Consonance and SE stands for Shannon Entropy. Thus, the order score is represented by the mean of the first three features presented above and the complexity is represented by the Shannon Entropy.

For feature extraction music21 (http://web.mit.edu/music21/) was used, a Python library designed for computer-aided musicology, music analysis, and music composition. It provides a range of tools and functionalities for working with music data in a digital environment. Music21 enables researchers, composers, and music enthusiasts to explore, manipulate and analyze musical content programmatically.

Besides this toolkit, JSymbolic [10] was used to extract the pitch and rhythmic value skewness values for each song. jSymbolic is a Java-based software toolkit for analyzing symbolic music data. It offers a range of algorithms to extract and analyze musical features, providing insights into rhythm, melody, harmony, and composition structure.

Interval Harmony

In music theory, pitch difference between two distinct notes is referred to as an interval. Specifically, the interval encompassing 12 semitones is denoted as an octave, holding significance within musical contexts. Supplementary materials offer a classification system for intervals, dividing them into five distinct categories, i.e., perfect, major, minor, augmented and diminished intervals. Multiple mathematics

and physics research papers have demonstrated that when two sound frequencies maintain a straightforward integer ratio, their harmonious combination is perceived as more aesthetically pleasing. Hence, the formula proposed by Jin et. al. [10] was employed:

$$Interval\ Harmony\ = \sum_{i\ =\ 1}^{12} w_i * rti_i$$

Where w_i represents the weight of that interval measured in cents and rti_i is the ratio between that interval and the total interval present in that song

Symmetry

For this feature, jSymbolic [10] was used to extract the pitch and rhythmic value skewness. Afterwards, the absolute values of these two features were used to derive the following formula:

$$Asymmetry\ =\ |PS|\ +\ |RVS|$$

Where PS and RVS are the skewness values for pitch and rhythm.

JSymbolic skewness represents a value for how much the pitch and rhythm distributions are asymmetrical to the left or the right of the mean value. The absolute values were used because in this study, the skewness's sign had no significance since the data was skewed anyways.

After performing the calculation for asymmetry, the definition of symmetry was based on its complementary nature. This definition was applied following the normalization process.

Interval Consonance

For computing this feature, it was necessary to determine an approach for considering the absence of a linear relation between the weight of an interval and how well it sounds based on the ratio between the sound frequencies.

For achieving this, a theory proposed by Paul Erlich [12] was used to measure the consonance or dissonance of an interval, based on its weight in cents

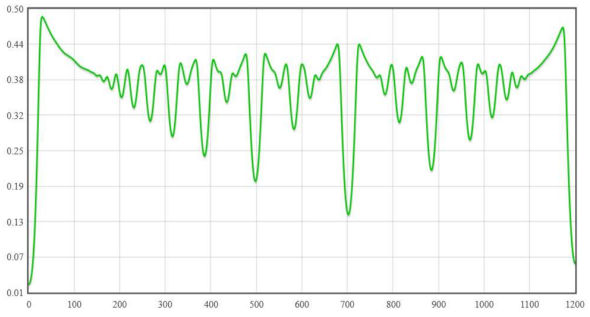

Fig. 1: Relationship between an interval's weight and its dissonance

The interval consonance score was computed by extracting the consonance of each interval present and then dividing the final consonance by the length of the song in seconds.

This division has the purpose of not giving a song a higher score solely because it has more intervals and therefore has more chances of having consonant intervals such as perfect fifths or perfect fourths that are highly consonant.

Shannon Entropy

Consider a finite set Ω, and let X be a random variable defined on this set. The distribution of X can be represented by the probability function p(x), denoting the probability of X taking the value x from the set Ω. Hence, the Shannon entropy, H(x), of a random variable X, representing a musical feature is defined as:

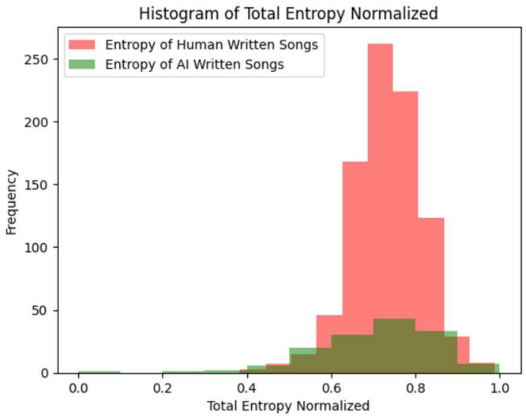

Fig. 2: Shannon Entropy histogram

$$H(x)\ =\ -\sum_{x\in\Omega} p(x)\ *\ log\ p(x)$$

The Shannon entropy serves as a metric for quantifying the average level of uncertainty inherent in a random variable X. This measure finds extensive application in assessing the level of disorder or chaos within the internal state of a system. To compute the entropy of music, it becomes essential to acquire the histogram of music attributes. For the final value for Shannon Entropy used as our complexity measure, entropies for both pitch and rhythm were averaged.

RESULTS

After using the methods that were described above for extracting the four features necessary for computing a final aesthetic score from both the datasets, the first batch of numeric scores are depicted in Table 1.

Feature	Human written songs		AI written songs	
	Mean	Median	Mean	Median
Interval Harmony	0.4148	0.3887	0.3325	0.2746
Symmetry	0.5714	0.5823	0.5276	0.5452
Interval Consonance	0.8409	0.8476	0.6918	0.7758
Shannon Entropy	0.7347	0.7349	0.7110	0.7331

Table 1: The mean and median values for the four extracted features for the AI and human written songs

Based on the observations presented in Table 1 and Figure 2, the two datasets exhibit relatively minor disparities in terms of entropy. This phenomenon can be attributed to the functioning principles of the music generation algorithms. Broadly speaking, these algorithms predict which musical fragment is most likely to follow the generated output up to the calculation point. Thus, one would expect them to not be significantly more unpredictable than human composers, considering their exposure to a diverse range of musical styles and their ability to emulate the spontaneity of a human composer.

The notable differences between the two datasets emerge in the metrics pertaining to order, particularly in the context of interval harmony and consonance. On average, compositions authored by human composers exhibit higher scores by 24.75% for interval harmony, 8.31% for symmetry, and 21.55% for interval consonance. These discrepancies arise from the fact that current artificial intelligence and computational models are incapable of comprehending the aesthetics underlying a musical piece. For instance, a computer lacks the ability to understand why a difference of 6 semitones may sound dissonant to a listener, while differences of 5 or 7 semitones are perceived as pleasant. These models are solely capable of mathematically and statistically reproducing what they have learned from their training dataset, but they lack the functional understanding of what makes a melody enjoyable.

Figure 3 provides a clearer depiction of the lack of expressiveness exhibited by the music generators. The concept of melodic symmetry is relatively easy to grasp as it can be easily translated into numerical values that can be utilized by various machine learning algorithms. However, harmony and consonance are challenging to describe in a manner comprehensible to a computer. It is observable that artificial melodies receive relatively lower scores when it comes to interval consonance, a crucial aesthetic aspect. Manual compositions, on the other hand, cluster relatively closely in this three-dimensional space, characterized by strong consonance and harmonies. In contrast, the outputs of intelligent agents appear to span a broader range of values. Although many outcomes tend to align with the physical space of real melodies, the agents still have much to learn before they can adequately reproduce human musical intelligence.

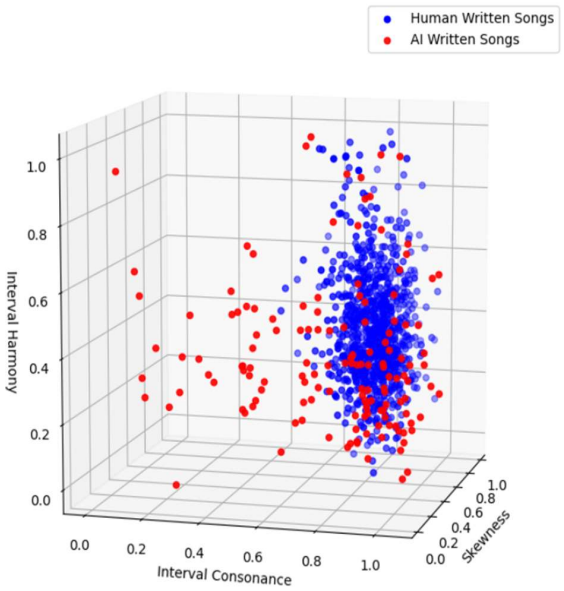

Fig. 3: The data set represented according to the 3 dimensions of order: Interval Harmony, Interval Consonance and Symmetry

After feature extraction, the scores were aggregated using Birkhoff's aesthetic score formula and the results are depicted in Table 2.

	Mean	Median	Standard Deviation
Human written songs	0.84173	0.82125	0.15781
AI written songs	0.75552	0.67687	0.28687

Table 2: Metadata extracted based on the Birkhoff score for the 2 datasets.

From Table 2 and Figure 4, may be seen that artificial intelligence (AI) has generated melodies that are comparatively weaker in terms of aesthetic quality when

compared to human-composed melodies. On average, the AI-generated songs scored 10.24%, with a nearly double standard deviation. This indicates that a wide range of aesthetically diverse melodies were generated, but the majority fell short in comparison to human-composed ones.

Figure 4 highlights that the AI system produced a few melodies with significantly high Birkhoff scores. Upon analyzing these results, it becomes apparent that the elevated scores are not attributable to the performance of the generation model, but rather the limitations of the Birkhoff formula.

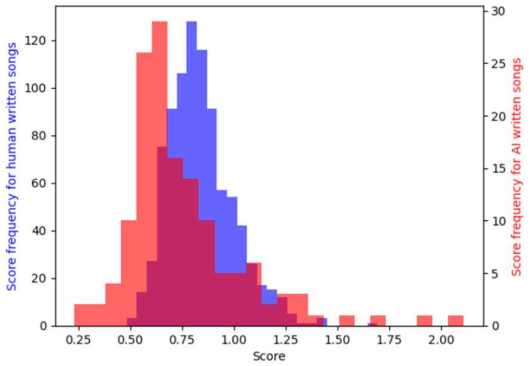

Fig. 4: Distribution of song scores obtained using Birkhoff's formula.

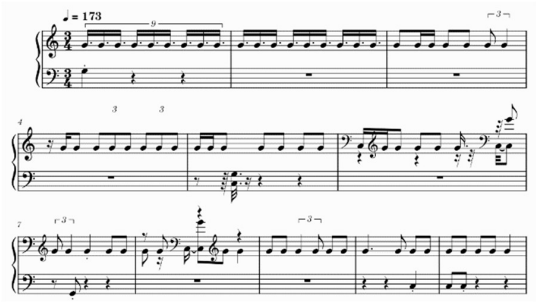

Fig. 5: Fragment of the musical piece generated by MuseNet when asked to create a classical piano piece in the style of Chopin

As depicted in Figure 5, the generator has not produced a composition of high aesthetic quality, instead, it represents a case where computational evaluation fails. This composition predominantly utilizes the note G, albeit with varying durations. This approach yields a high score since nearly all intervals have a size of 0 semitones (unison), resulting in perfect consonances, thereby obtaining a high score for interval consonance. Furthermore, this lack of creativity and emotion contributes to a low entropy score, as the majority of notes share the same pitch, and the rhythm lacks variation, leading to reduced complexity. These factors culminate in a significantly elevated final score for the composition, despite its lack of artistic value.

After manual analysis and correction of these type of corner cases, the resulting final aggregated scores are presented in Table 3.

	Mean	Median	Standard Deviation
Human written songs	0.84173	0.82125	0.15781
AI written songs	0.725	0.67272	0.22361

Table 3: Metadata extracted based on the Birkhoff score for the 2 datasets after correction.

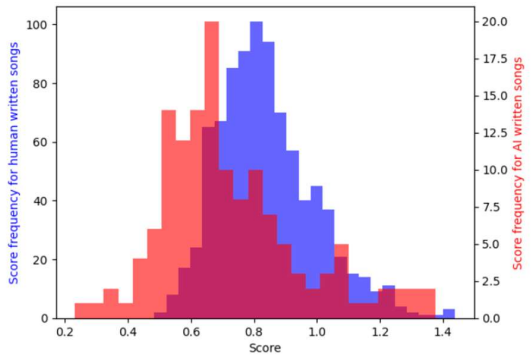

Fig. 6: Distribution of song scores obtained using Birkhoff's formula after correction.

From figure 6 and table 3, results resembling real-world scenarios can be observed. The average scores of the AI generated songs have decreased from 0.755 to 0.725, while the median value has decreased from 0.67687 to 0.67272. A significant change can be noticed in the standard deviation, which has decreased by 22%. This reduction indicates that the initial iteration of the calculation also considered points where the formula did not accurately capture the true aesthetic value of a melody.

CONCLUSIONS

This paper provides a way to measure a musical's piece aesthetic measure in a computational way and compares two song libraries: one written by human composers and the other one written by AI. It can be concluded that artificial intelligence has made exponential advancements in recent years, however, there is still a long way to go before it can consistently produce high-quality melodies on a large scale. As perceivable from figure 6, there are many artificially generated songs that obtain decent scores, but the majority hover around a score of 0.67. In contrast, compositions

written by composers achieve a score close to 0.82, highlighting a substantial difference between the two.

Despite machine learning algorithms being capable of generating compositions with a certain level of coherence and structure, they may encounter challenges in conveying deep emotions and profound sentiments that are characteristic of human interpretation. While artificially generated music can be technically pleasing, the lack of intuition, human experience, and sensitivity can lead to it being perceived as soulless and devoid of emotion. Authentic emotions and feelings in music are often conveyed through subtle interpretation of notes, tones, vocal expression, or instrumental performance. These elements can be difficult to accurately reproduce by generators, as they involve subjective aspects and human intuitions.

The results obtained in this paper represent a promising starting point for models that generate musical compositions. Although the studied pieces obtained lower scores on average compared to composed ones, among them, there are aesthetic compositions characterized by beautiful chords and harmonies. Considering that this field is still in its early stages, we believe that with the emergence of numerous datasets and studies, a model will soon arise capable of creating cohesive pieces across different musical styles based on both user-provided melody fragments and text inputs.

In the future, this project can be beneficial to researchers working in this domain who require a quantitative method to evaluate the qualitative aspects of a composition without relying on a group of human listeners to provide scores for each piece. This measure can subsequently be employed by various machine learning algorithms.

REFERENCES

1. Paroiu, R., & Trausan-Matu, S. (2023). Measurement of Music Aesthetics Using Deep Neural Networks and Dissonances. Information, 14(7), 358. MDPI AG. ISSN: 2078-2489, http://dx.doi.org/10.3390/info14070358
2. Epperson, G. (2022). "Music Art Form, Styles, Rhythm, & History". Encyclopaedia Britannica. Chicago: Encyclopædia Britannica, Inc.
3. Francesconi, E. (2022). “The winter, the summer and the summer dream of artificial intelligence in law”, Artif Intell Law 30, 147–161 https://doi.org/10.1007/s10506-022-09309-8
4. Newman, D. "Exploring The Ins And Outs Of The Generative AI Boom". Forbes.https://www.forbes.com/sites/danielnewman/2023/03/14/exploring-the-ins-and-outs-of-the-generative-ai-boom/ Retrieved on 14th March 2023
5. Trausan-Matu, S. & Dathan, B. (2016) Perceived aesthetics of user-modifiable layouts: a comparison between an unspecified design and a GUI, in A. Iftene & J. Vanderdonckt (Eds.), Romanian Conference on Human-Computer Interaction (RoCHI 2016), pp.22-25
6. Birkhoff, G.D. (1933). “Aesthetic Measure”, Harvard University Press: Cambridge, MA, USA.
7. Ziyu W., & Ke C., & Junyan J., & Yiyi Z., & Maoran X., & Shuqi D., & Xianbin G., & Gus X. (17th August 2020). “POP909: A Pop-Song Dataset For Music Arrangement Generation” arXiv:2008.07142
8. Hawthorne C., & Stasyuk A., & Roberts A., & Simon I., & Huang C. Z. A., & Dieleman S., & Elsen E., & Engel J., & Eck D. (2019). “Enabling factorized piano music modeling and generation with the MAESTRO dataset,” in 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA.
9. Payne, C. (25th April 2019). "MuseNet." OpenAI, openai.com/blog/musenet Retrieved on 20th June 2023
10. McKay C., & Fujinaga I. (2013). “jSymbolic: A Feature Extractor for MIDI Files”, International Conference on Mathematics and Computing.
11. Jin, X. & Zhou, W. & Wang, J. & Xu, D. & Rong, Y. & Cui, S.. (2023). “An Order-Complexity Model for Aesthetic Quality Assessment of Symbolic Homophony Music Scores”. 10.48550/arXiv.2301.05908.
12. Erlich P. (1998). “Tuning, Tonality, and Twenty-Two-Tone Temperament”, Xenharmonikôn journal.