An Empirical Study on How People Perceive AI-generated Music

Hyeshin Chu
hyeshinchu@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Joohee Kim
joohee@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Seongouk Kim
zjqvp@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Hongkyu Lim
limhongkyu1219@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Hyunwook Lee
gusdnr0916@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Seungmin Jin
skyjin@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Jongeun Lee
jongeunlee@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Taehwan Kim
taehwankim@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Sungahn Ko*
sako@unist.ac.kr
UNIST
Ulsan, Republic of Korea

ABSTRACT

Music creation is difficult because one must express one's creativity while following strict rules. The advancement of deep learning technologies has diversified the methods to automate complex processes and express creativity in music composition. However, prior research has not paid much attention to exploring the audiences' subjective satisfaction to improve music generation models. In this paper, we evaluate human satisfaction with the state-of-the-art automatic symbolic music generation models using deep learning. In doing so, we define a taxonomy for music generation models and suggest nine subjective evaluation metrics. Through an evaluation study, we obtained more than 700 evaluations from 100 participants, using the suggested metrics. Our evaluation study reveals that the token representation method and models' characteristics affect subjective satisfaction. Through our qualitative analysis, we deepen our understanding of AI-generated music and suggested evaluation metrics. Lastly, we present lessons learned and discuss future research directions of deep learning models for music creation.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; • **Computing methodologies** → *Neural networks*; • **Human-centered computing** → **Empirical studies in HCI**; **User studies**.

KEYWORDS

human-AI interaction, symbolic music generation, deep learning, subjective evaluation

*Corresponding Author

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557235

ACM Reference Format:
Hyeshin Chu, Joohee Kim, Seongouk Kim, Hongkyu Lim, Hyunwook Lee, Seungmin Jin, Jongeun Lee, Taehwan Kim, and Sungahn Ko. 2022. An Empirical Study on How People Perceive AI-generated Music. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3511808.3557235

1 INTRODUCTION

With the tremendous advancement of social media, the ability to produce attractive content has become significant for both professional and novice content creators. Content creators should consider not only their interests but also the preferences and likes of others to allure more viewers. In addition to music-focused platforms [26], video-focused services, such as TikTok [3], have fascinated a huge number of users through new content types that combine video and music. However, it is difficult to compose music that can satisfy both content creators and the audience [34]. Since professional music composition requires a great deal of skill and strong background knowledge, novice users can barely make enjoyable musical content [62]. Social media platforms provide some features, including a large variety of commercial music, to assist the users in selecting a song best suited to their needs. However, it is time-consuming and expensive to earn the right to use the music [55].

AI-based music generation can be an alternative because it makes the music composition process much easier and more convenient for both professional and novice content creators [40]. With the advancement of neural network architectures and the increased interests in the role of AI in enhancing creative human activities, more studies have focused on novel approaches to generating music with AI, including autonomous music generation models [9, 14, 22], and human-AI co-creation systems for music [19, 40, 41, 58].

Current AI music generation research has achieved great success in the algorithmic aspect [27, 67, 77] and has diversified in other aspects, including tasks and type of generated music [4–6, 25]. For instance, several studies have proposed novel frameworks or improved model performance in terms of objective metrics, such as efficiency or accuracy [29, 46]. However, few studies have discussed the significance of understanding the audience' subjective

satisfaction with AI-generated music [24]. Human satisfaction and individual subjectivity are key components in determining the value of music, as in the case of other types of artwork [21].

In this paper, we aim to explore how novice users perceive the music created by the state-of-the-art symbolic music generation models. We set our target users as novice users without professional musical knowledge who have a great interests in music, since the aforementioned services aim to inspire novice users to express their creativity [60, 71]. To this end, we conduct the following tasks. As AI music generation models have diversified in various aspects, including generation tasks, type of generated music, and backbone architecture [4, 25], we first survey previous symbolic music generation models and subjective evaluation criteria. We then derive nine metrics for evaluating AI music based on the collected subjective criteria, including *overall*, *creativity*, *naturalness*, *melodiousness*, *richness*, *rhythmicity*, *correctness*, *structureness*, and *coherence*. Second, we conduct a comprehensive experiment, where 100 participants recruited online listen to the songs generated by state-of-the-art music generation models (e.g., Music Transformer, Transformer-GANs) with the MAESTRO dataset [23] and evaluate the songs based on the derived metrics. We reproduce both conditional and unconditional types of models for the experiment. The conditional type means that a model needs an input song, while unconditional type models do not need such input songs for music generation. The experiment results indicate that human satisfaction with each model has significant difference in several metrics, including *overall*. Also, participants consider *melodiousness* the most important, among the metrics. We also find that token representations and models' characteristics affect the subjective perception of novice users. The qualitative analysis results show that AI-generated music enriches user experience by inspiring creative activities, complementing the experiences using other media (e.g., video), and stimulating human emotions, as real music does. Finally, we present the lessons learned and implications for future studies on music generation models and human satisfaction towards them. To our knowledge, this is the first attempt to investigate how human perceive AI-generated music with comprehensive experiments and subjective metrics.

The main contributions of this work include the following:

- Deriving the taxonomy on music generation models based on voice, texture, task, and architectures,
- Developing of subjective metrics for AI music evaluation,
- Performing a comprehensive user study with four state-of-the-art music generation models and 100 online participants,
- Lessons learned and discussion on future research directions.

In the next section, we introduce prior works related to our research. In section 3, we characterize music generation models based on the taxonomy of music. In section 4, we describe our subjective evaluation metrics and baseline models. In section 5 and 6, we illustrate our user study and the result. Lastly, we represent the lessons we learned regarding AI music generation models and their subjective evaluation metrics and discuss future research direction.

2 RELATED WORK

In this section, we introduce some prior works that discuss music generation models and evaluation metrics.

2.1 Music Generation Models

Recent deep learning music generation models have diversified in many respects. One standard that differentiates them is the level of autonomy [4]. Some models are fully autonomous and require no human intervention in the music composition process. For instance, Dong et al. [14] proposed a fully autonomous music generation framework that creates multi-track music from scratch. There are also various human-AI co-creation methods that help people create music. Some studies have addressed how music generation systems interact with the users. For example, Suh et al. [58] found that AI played a role in the social dynamics among human composers. Frid et al. [19] proposed a user interface that enhances the interactivity between AI and human composers. Louie et al. [40] pointed out that users have different level of ability to create music, and present AI-steering tools for novices.

Another important standard by which to characterize music generation models is token representation [4, 5]. Models use either audio or symbolic representations for their input and output. While audio-representation models deal with continuous variables, symbolic music generation models handle discrete variables [4]. In addition, symbolic music generation models can be divided into two types: image-based and language-based. Image-based methods include representations such as MIDI and pianoroll [15]. In contrast, language-based approaches contain representations such as MIDI-like event [29], REMI [31], and CP [27].

Music generation models have also developed with the advancement of diverse neural network architectures. In particular, various music generation models are based on recurrent neural network (RNN) [8, 22, 75], long short-term memory (LSTM) [18, 43, 56], autoencoder [38, 59, 74], transformer [7, 29, 37], and generative adversarial network (GAN) [35, 51, 77]. Although several studies have proposed novel music generation models, there is potential to further discuss human satisfaction with AI-generated music and improve models regarding the satisfaction.

2.2 Evaluation Metrics

Several studies have discussed evaluation methods for AI-generated music. Although various studies have highlighted the significance of subjective evaluation in music generation, they have also mentioned the difficulty of evaluating subjective satisfaction through listening tests because such tests are time-consuming and the experimental setting can be affected by several variables (e.g., participants' characteristics, questions phrasing, etc.) [4, 5, 70]. For these reasons, prior research on the evaluation of AI-generated music has paid more attention to **objective evaluation** than subjective evaluation [50, 70]. In particular, the objective metrics include pitch-related and rhythm-related ones [13, 45, 50, 67].

However, the objective evaluation focuses on model-centric accuracy, without considering human satisfaction with the generated output. Therefore, a growing number of studies emphasize the significance of the **subjective evaluation** of AI-generated music [24]. The Turing test, one of the earliest methods for evaluating human satisfaction with machine-generated music, asks people to identify whether the music was generated by a human or a machine [2, 22, 28, 30, 37, 54]. Various other types of listening tests including preference questions have been developed.

Moreover, some studies have suggested detailed subjective evaluation metrics. For example, Wu et al. [67] proposed four metrics (overall quality, impression, structureness, and richness). However, they compared AI-generated songs with human-composed ones without comparing the songs generated using different AI models. Carnovalini et al. [6] suggested creativity as another criterion, but the study did not contain experimental designs or evaluation results. Dong et al. [14] suggested subjective evaluation metrics (harmony, rhythm, musical structure, coherence, and overall rating), but that study evaluated only a proposed model, without comparing it to the generated music pieces generated using other models. Furthermore, no prior research has asked participants' opinions on the AI-generated music or explored human perception with the music. Therefore, our research suggests subjective evaluation metrics, compare AI-generated songs created using various models, and discuss the participants' opinions of the generated music and evaluation criteria. Through these efforts, we deepen our understanding of AI-generated music and subjective evaluation criteria.

3 PRELIMINARY STUDY

To understand the characteristics of music generation models and select the ones for our experiment, we take the following steps.

3.1 Characterizing Music Generation Models

To begin with, we review several surveys on music generation and list up existing music generation models. Studying prior surveys, we find that they used various standard to classify the models and they do not contain experiment result to compare the model performance or human perception with the generated music of each model. In order to address this issue, we delineate the taxonomy of AI music generation models based on the number of voices, musical texture, generation task, and architecture, to select the baseline models for our experiment.

As a result, we categorize 40 music generation models based on these criteria, as shown in Table 2. Firstly, we classify the models into two groups depending on the number of voices in generated music [44]:**single-voice** and **multi-voice**. Single-voice music refers to music played with a single instrument, while multi-track music refers to music pieces that have a multi-track [4]. Among the 40 models studied, 23 of them generate single-voice music, 16 models propose multi-voice music, and one provides both. Second, there are four **musical textures**: *monophony* , *polyphony* , *homophony* , and *heterophony*. All the studies focus on generating monophonic or polyphonic music. These music generation models can then be divided into two groups based on the **generation task** of each model (i.e., conditional and unconditional music generation). *Conditional* music generation models uses a musical fragment (e.g., a musical theme) as an input to condition the generative [53]. On the other hand, *unconditional* music generation models create music from scratch [27]. Each model conducts conditional or unconditional music generation tasks, and some research has proposed models for both tasks. Various neural network **architectures** are used as the backbones of music generation models. The 40 models that we introduce are based on either one of the recurrent neural network (RNN), long short-term memory (LSTM), variational auto-encoder (VAE), transformer, and generative adversarial network (GAN).

3.2 Selecting Models for the Experiment

Table 1 represents the baseline models for our experiment. Among the 40 models, we first exclude the ones that generate only monophonic music because most of the 40 models generate *polyphonic* music. Also, we focus on *single-voice* music because we aim to understand the human perception with AI-generated music rather than explore various features of AI music composition models, such as accompaniment generation. In the same context, we exclude some of conditional music generation models that utilize anything other than a short music piece as an input, such as emotion class [32]. Then, we select the state-of-the-art models among them. To evaluate the human satisfaction with both *unconditional* and *conditional* music generation models, we select three models for each task. Consequently, we finalize the AI music generation models for our experiment, as shown in Table 1.

4 METHOD

In this section, we introduce the dataset, baseline models, and subjective evaluation metrics for our experiment. In particular, we highlight the major differences between music generation models. Also, we address how we categorize and suggest our subjective evaluation metrics by reviewing 40 music generation models.

4.1 Dataset and Models

In this experiment, we use the MIDI and Audio Edited for Synchronous TRacks and Organization(MAESTRO) dataset [23]. It contains about 200 hours of piano performances recorded from the International Piano-e-Competition. Additionally, MAESTRO consists of paired audio and MIDI files and the metadata files for each pair. The metadata includes information about the composer, title, year of performance, and duration of each music piece. The MAESTRO dataset has been utilized in several music generation studies [10, 46, 63].

As shown in Table 1, we select three models for *unconditional* and *conditional* music generation task, respectively. For unconditional music generation task, we use Music Transformer [29], Compound Word Transformer [27], and Transformer-GANs [46] as baseline model. For conditional music generation task, we use Music Transformer [29], Theme Transformer [53], and Transformer-GANs [46]. In the evaluation study, we provide songs generated from the following baseline models to compare human perception with music generated by each model. As shown in Table 1, the baseline models have major differences in terms of their tasks, token representation methods, and characteristics in model.

4.2 Evaluation Metrics

Table 3 represents how we categorize subjective evaluation metrics of the 40 music generation models. First, we collect all subjective evaluation metrics that have been used in the 40 music generation models. We investigate each metric and its definitions to categorize the metrics. Among the 40 models, 10 did not conduct any listening tests to evaluate users' subjective satisfaction [17, 33, 42, 48, 49, 56, 57, 61, 65, 72]. Hadjeres et al. [22] only conducts the Turing test without evaluating the songs based on specific metrics. Muhamed et al. [46] includes a ranking type question that asks the participants to order the music pieces based on their preference. The other studies evaluate human satisfaction based on specific subjective evaluation

Table 1: Baseline Models

Model	Music Transformer	Compound Word Transformer	Transformer-GANs	Theme Transformer
Task	Unconditional, Conditional	Unconditional	Unconditional, Conditional	Conditional
Token Representation	MIDI-like, event-based [47]	Compound Words (CP)	MIDI-like, event-based [47]	Note, Metric, Theme-related tokens
Model Characteristics	Relative-attention	Transformer decoder for different types of tokens	Adversarial losses for long-term coherence	Theme-conditioned Transformer

Table 2: Music Generation Models

Voice	Texture	Task	Architecture	Model
Single	Monophony	Unconditional	RNN	[65]
			VAE	[52]
			GAN	[35]
		Conditional	GAN	[69]
	Polyphony	Both	Transformer	[27, 29]
			GAN	[46]
		Unconditional	RNN	[22]
			LSTM	[56]
			VAE	[66]
			Transformer	[7, 72, 73]
		Conditional	RNN	[8]
			LSTM	[9, 18, 43]
			VAE	[59, 74]
			Transformer	[31, 68]
				[32, 49, 53]
Multi	Polyphony	Both	GAN	[14, 39]
		Unconditional	VAE	[38]
			Autoencoder	[1, 61]
			Transformer	[12, 67, 73]
			GAN	[16]
		Conditional	RNN	[75]
			VAE	[42, 57]
			Transformer	[17, 33, 48]
			GAN	[51, 77]

metrics. In total, there are 39 unique metrics. We categorize the metrics into 14 groups based on their definitions, as shown in Table 3. We exclude theme-specific (e.g., theme repetition, theme timing [53]) or emotion-related metrics (e.g., valence, arousal [18, 32]). Additionally, we rule out the metrics that appeared only once (e.g., singability [77]). Then, we add one criterion(*creativity*), since several prior research emphasize the role of AI to boost human creativity in music composition [6, 41]. Consequently, we have nine subjective evaluation metrics and their definitions as follows: For unconditional music generation models, we utilize eight criteria: *overall*, *creativity*, *naturalness*, *melodiousness*, *richness*, *rhythmicity*, and *structureness*. For conditional music generation models, we use the eight criteria above in addition to a ninth criterion: *coherence*. For both tasks, we utilize 7-point Likert scale.

- **Overall**: What is your overall satisfaction with the music?
- **Creativity**: Is the music piece novel, valuable, and original?
- **Naturalness**: Does the piece sound like an expressive human performances?
- **Melodiousness**: How musical and harmonious is the piece?
- **Richness**: How diverse and interesting is the piece?
- **Rhythmicity**: Does the music have a unified rhythm?
- **Correctness**: Does the music play with any technical glitch (e.g., a sudden pause)?
- **Structureness**: Are there any structured patterns, such as repeating themes or the development of musical ideas?
- **Coherence**: Is the conditionally generated piece similar to the reference?

We have two goals regarding these subjective evaluation metrics. First, we aim to use metrics that help us to evaluate various elements of music. To this end, we include not only the metrics for music itself(i.e., *overall*, *creativity*, *melodiousness*, *richness*, *rhythmicity*, and *structureness*) but also the metrics only for AI-generated music(e.g., *naturalness*, *correctness*, and *coherence*). Second, our user study aims to evaluate the satisfaction of novice users without professional experience or education in music. Therefore, we intend to provide clear definitions that are easy for novice users to understand.

5 USER STUDY

To evaluate human perception with the generated music pieces, we conducted a user study of 100 participants. The user study aims to understand how people perceive the generated songs based on the suggested criteria. As shown in Fig. 1, there were two separate tests in the experiment, one to evaluate human satisfaction with unconditional music generation, and the other for conditional music generation. We used eight criteria for unconditional music generation tests and nine for conditional music generation tests.

5.1 Study Design and Procedure

For the user study, we created a questionnaire via Google Forms. As shown in Fig. 1, we used three models for unconditional and conditional music generation test, respectively. To select the music clips for the user study, we first generated 10 songs using each model. Then, we selected the three most satisfying songs from each model. Before conducting the user study, we conducted an internal test with three of the authors to check whether the questionnaire is clear to novice users without professional knowledge or experience in music. Also, to ensure that the overall test is completed within an hour, thereby preventing auditory fatigue in the participants, we provided three songs from each model. We provided full compensation, $9.96 per hour, which is above current hourly minimum wage($7.25 [64]), to every participant who finished the user study.

Unconditional music generation test provide nine music pieces because they do not need theme (short music pieces) to generate the music. On the other side, conditional music generation test include nine music pieces and nine theme, because they need theme to generate the music. As we trained the models by MAESTRO dataset [23], the generated songs are piano music. Each music piece is between 40 seconds and 1 minute in length, long enough that participants can feel the music and short enough to prevent auditory fatigue. Fig. 1 summarizes the data collected from the user study.

Fig. 2 represents the procedure of our user study. As shown in Fig. 2, the user study includes three steps: *demographic survey* to gather participants' characteristics, a *listening test* to evaluate the generated songs based on our proposed criteria, and a *post survey* to understand participants' opinions on the criteria. First, participants read the goal of the user study: to gain an understanding of human satisfaction with AI-generated songs. Then, they read the restrictions, compensation details, tutorial, and our institution's ethics guidelines. We informed that the songs are generated by

Table 3: Categorizing Subjective Evaluation Metrics

Category	Metrics	Definition	Models
Overall	Overall	N/A	[14, 16, 27, 32, 39, 46, 53, 67, 73]
		Which music piece has the best overall musicality?	[66, 74]
Melodiousness	Melody	N/A	[1, 38, 59, 68]
	Musical	Which they thought was more musical?	[29, 52]
		Are the music notes' relationships natural and harmonious?	[77]
	Harmonious	N/A	[1, 16, 29, 39, 51, 59, 68]
		Does the music clip have pleasant harmony?	[14]
Naturalness	Naturalness	N/A	[59]
	Real	How realistic is the sequence?	[35, 69]
	Humanness	Does the piece sound like expressive human performances?	[27]
		How well it sounds like a piece played by human?	[32]
Coherence	Coherence	Is the music clip coherent?	[14]
	Fidelity	Is the conditionally generated piece similar to the reference, from which the condition lead sheet was taken from?	[27]
Correctness	Correctness	Perceived absence of composing or playing mistakes?	[27]
	Integrity	N/A	[38]
Structuredness	Structureness	N/A	[38, 53, 67, 73]
		Whether there are structural patterns such as repeating themes or development of musical idea	[27]
		Does the accompaniment flow dynamically with the structure of the melody?	[74]
		Does the music clip have clear musical structure?	[14]
		To what extent the music sample exhibits an organizational structure?	[75]
Rhythmicity	Rhythmicity	N/A	[1, 16, 38, 39, 59, 68, 73]
		Does the music clip have unified rhythm?	[14]
		Does the music sound fluent and pause suitably?	[77]
Richness	Interesting	How interesting is the song?	[32, 35, 69]
		Diversity and interestingness	[27, 67]

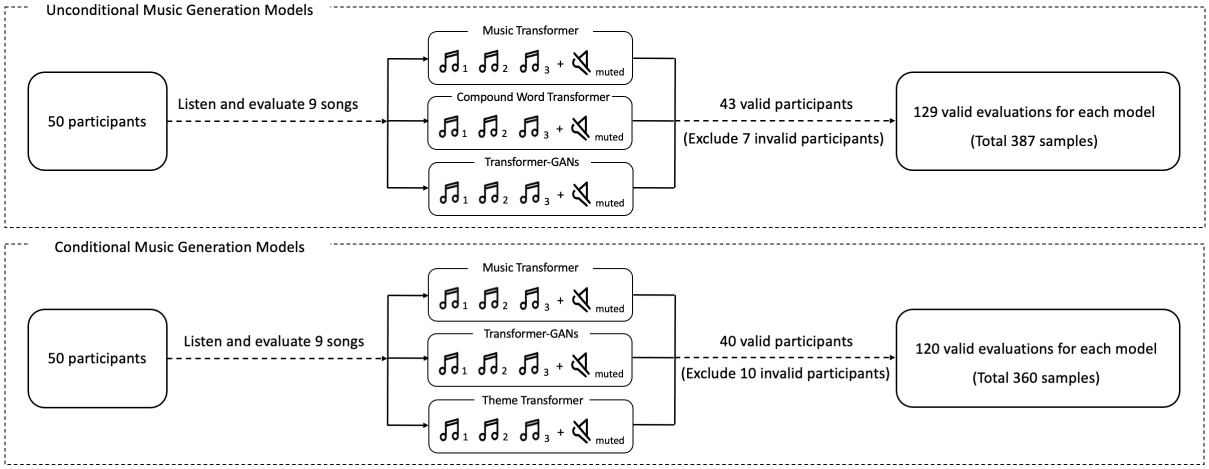

Figure 1: User Study and Data Collection

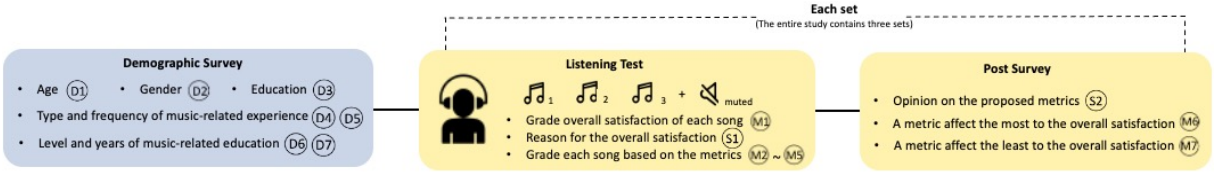

Figure 2: The Procedure of User Study

AI and participants need to listen to the full music piece from the beginning to the end of each song. We noted that the questionnaire has a blank audio clip for each question that contained no sound to prevent insincere answers. We exclude any evaluation that contains insincere answer (e.g., giving a score to muted clip), and it was announced to the participants before they start the user study.

Consenting to the aforementioned restrictions, compensation details, tutorial, and the ethics guideline, the participants move on to demographics survey. As shown in Fig. 2, the demographics survey contain seven multiple-choice questions, including their age (D1), gender (D2), education (D3), interest in music (D4–D5), and level of background knowledge of music (D6–D7).

Next, the participants move on to the listening test. As shown in Fig. 1, we designed two separate questionnaires, one each for unconditional and conditional music generation. A participant answer to either questionnaire. As shown in Fig. 2, the listening test consists of three sets. In each set, participants first listen to four songs and mark their overall satisfaction with each song (M1). Next, the participants state the reason for their score in minimum of 100 words (S1). They were instructed to share their reasons for each

Table 4: Song comparison of Unconditional and Conditional Music Generation Models

	Model	Music Transformer				Compound Word Transformer				Transformer-GANs			
	Metric	Song 1	Song 2	Song 3	p-value	Song 1	Song 2	Song 3	p-value	Song 1	Song 2	Song 3	p-value
(a) Unconditional	Overall	4.883 ± 1.617	3.674 ± 1.307	5 ± 1.447	0.735	4.791 ± 1.250	5.303 ± 1.249	4.977 ± 1.229	0.497	3.209 ± 1.824	2.558 ± 1.403	3.302 ± 1.862	0.806
	Creativity	4.163 ± 1.711	3.791 ± 1.608	4.674 ± 1.639	0.163	4.023 ± 1.607	4.791 ± 1.536	4.244 ± 1.428	0.539	3.883 ± 1.943	3.302 ± 1.622	3.884 ± 1.820	1
	Naturalness	5.163 ± 1.613	3.744 ± 1.630	5 ± 1.479	0.659	4.791 ± 1.322	5.512 ± 1.301	4.953 ± 1.219	0.592	3.070 ± 1.922	2.628 ± 1.599	3.372 ± 1.779	0.439
	Melodiousness	5.023 ± 1.811	4 ± 1.510	5.116 ± 1.401	0.797	4.791 ± 1.322	5.814 ± 1.167	5.256 ± 1.241	0.102	3.419 ± 1.890	2.767 ± 1.612	3.512 ± 1.703	0.809
	Richness	4.488 ± 1.690	3.535 ± 1.515	4.628 ± 1.599	0.702	4.323 ± 1.667	5.349 ± 1.310	4.395 ± 1.349	0.833	3.791 ± 1.824	3.047 ± 1.540	3.837 ± 1.697	0.901
	Rhythmicity	5.023 ± 1.811	3.977 ± 1.548	5.047 ± 1.293	0.945	4.698 ± 1.593	5.628 ± 1.239	5 ± 1.078	0.312	3.465 ± 1.676	3.047 ± 1.698	3.256 ± 1.831	0.581
	Correctness	5.163 ± 1.816	4.209 ± 1.719	4.884 ± 1.687	0.472	5.023 ± 1.635	5.302 ± 1.607	5.256 ± 1.511	0.501	3.512 ± 1.945	3.395 ± 1.857	3.651 ± 1.915	0.737
	Structureness	4.581 ± 1.701	4.093 ± 1.428	5.140 ± 1.340	0.098	4.628 ± 1.525	5.302 ± 1.267	4.860 ± 1.357	0.450	3.047 ± 1.670	3.162 ± 1.791	3.489 ± 1.717	0.243
	Model	Music Transformer				Transformer-GANs				Theme Transformer			
	Metric	Song 1	Song 2	Song 3	p-value	Song 1	Song 2	Song 3	p-value	Song 1	Song 2	Song 3	p-value
(b) Conditional	Overall	3.675 ± 1.439	4.525 ± 1.533	3.7 ± 1.503	0.942	4.075 ± 1.679	2.725 ± 1.533	3.55 ± 1.413	0.213	5.325 ± 1.292	4.975 ± 1.541	5.7 ± 1.030	0.155
	Creativity	3.85 ± 1.441	5.15 ± 1.526	4.25 ± 1.392	0.252	4.05 ± 1.774	3.975 ± 1.725	3.775 ± 1.620	0.651	5.05 ± 1.303	4.7 ± 1.536	5.2 ± 1.520	0.476
	Naturalness	3.775 ± 1.589	5.075 ± 1.233	4.05 ± 1.612	0.443	4.425 ± 1.745	3.025 ± 1.537	3.65 ± 1.652	0.645	5.125 ± 1.503	4.9 ± 1.530	5.275 ± 1.265	**0.048**
	Melodiousness	3.925 ± 1.618	5.025 ± 1.508	4.125 ± 1.806	0.605	4.475 ± 1.732	3.375 ± 1.713	3.75 ± 1.639	0.226	5.25 ± 1.356	5.45 ± 1.413	5.625 ± 1.336	0.065
	Richness	3.9 ± 1.7	4.8 ± 1.676	4.175 ± 1.412	0.458	3.85 ± 1.476	3.975 ± 1.851	3.9 ± 1.530	0.198	4.775 ± 1.475	5 ± 1.581	5.225 ± 1.573	0.891
	Rhythmicity	3.55 ± 1.923	4.9 ± 1.513	3.525 ± 1.565	0.950	4.075 ± 1.794	3.475 ± 1.673	3.825 ± 1.626	0.261	5.3 ± 1.327	5.475 ± 1.284	5.625 ± 1.218	0.518
	Correctness	3.487 ± 1.816	4.975 ± 1.423	4.2 ± 1.676	0.065	4.475 ± 1.612	3.75 ± 1.813	4 ± 1.658	0.657	5.05 ± 1.564	5.275 ± 1.499	5.2 ± 1.418	0.221
	Structureness	3.975 ± 1.768	4.95 ± 1.532	3.875 ± 1.452	0.790	4.5 ± 1.466	3.4 ± 1.562	3.775 ± 1.508	0.305	5.25 ± 1.337	5.325 ± 1.421	5.55 ± 1.094	**0.04**
	Coherence	4.225 ± 1.796	5.3 ± 1.363	4.025 ± 1.651	0.604	4.65 ± 1.681	3.775 ± 1.943	4.15 ± 1.783	0.479	5.325 ± 1.367	5.35 ± 1.130	5.525 ± 1.248	0.227

evaluation, including the highest and lowest scores. Then, participants listen to the four songs again and evaluate their satisfaction based on our suggested criteria (M2 – M5). The definition of each evaluation criteria were provided to the participants. We use 7-point Likert scale for every question in the listening test. As shown in Fig. 2, participants answer questions in all three sets, each of which contain music generated by one model. Overall, each participant listen to and evaluate their satisfaction with nine songs generated using three models, by answering to 15 multiple choice questions and three short answer question. After finishing the listening test, the participants move on to the post survey.

The post survey aims to increase our understanding of participants' opinions on the evaluation criteria. As Fig. 2 indicates, participants answer to a short answer question and two multiple choice questions in the post survey. The short answer question (S2) asks participants to state their opinions of the criteria or suggest their own criteria for AI-generated music in minimum 100 words. It is designed to identify whether participants think the suggested criteria are sufficient to evaluate their satisfaction with the generated songs. The two multiple choice questions ask participants mark the most effective criterion (M6) or the least effective one (M7) to *overall*. In summary, participants answer to 21 multiple choice questions and six short answer questions during listening test and post survey.

5.2 Participants

We recruited 100 participants online (Prolific) for the experiment (Fig. 1). Since the experiment has two separate tests, *unconditional* and *conditional* music generation, 50 participants took each test. The participants were between the ages of 18 and 64 (M=26.14, SD=5.38, 42 female, 41 male). Among the 83 participants who provided sincere answers, none of them are professional musician. 26 of them have never learned how to play musical instruments, 47 participants have learned how to play at least one musical instruments, and 10 participants have learned how it from a professional musician. In addition, 82 out of 83 participants have experienced some type of music-related activities including listening to music (N=71), attending music concert or festival (N=47), playing musical instruments (N=29), and creating music (N=5), while multiple selections were available. In consequence, we found that the participants are novice users with little knowledge in music, but at the same time, they have great interests in music.

6 RESULT

In this section, we present the result of quantitative and qualitative analysis of the user study. The results show that the characteristics of each model affect to human satisfaction with each metrics. For qualitative analysis, three authors reviewed all open-ended text responses and identified the evaluation metrics through the iterative discussion process.

There were no significant differences among the three songs from each model. To ensure a fair comparison among the three models, we checked whether the three songs generated using each model differed significantly regarding participants' satisfaction. To this end, we asked the participants to indicate their satisfaction with every metric for every song in the evaluation study. Table 4 (a) shows the song comparison result of **unconditional music generation** models. As Table 4 (a) shows, the three songs generated from Music Transformer do not differ significantly from each other in any of the eight metrics, including *overall* (p=0.735). Likewise, the three songs generated using Compound Word Transformer (p=0.498) and Transformer-GANs (p=0.806) also do not differ significantly in any of the metrics including *overall*. Therefore, for each model, the three songs generated do not differ significantly regarding any metric. Table 4 (b) shows the differences among the three songs generated by each **conditional music generation** models. Table 4 (b) shows that neither the three songs generated by Music Transformer (p=0.943), nor the three generated by Transformer-GANs (p=0.213) differ significantly from each other in any of the nine metrics. In contrast, the three songs generated by Theme Transformer have significant difference in *naturalness* (p=0.049) and *structureness* (p=0.040), but not in *overall* (p=0.155). Based on these results, to compare human perception towards three models, we used the evaluations of all three songs generated by each model rather than choosing one song from each model.

Highlighting the co-occurrence relationship between different types of tokens enriches the melody. Compound Word Transformer is distinct from the other models because it considers different types of tokens, thereby better reflecting the role of each token type in music generation. As Figure 3 (a) shows, our results support the strength of this model in this regard. While one-way ANOVA test showed that the three models differ significantly from each other in terms of *melodiousness*, Tukey post-hoc analysis found that Compound Word Transformer (μ=5.287, SD=1.313) has significantly

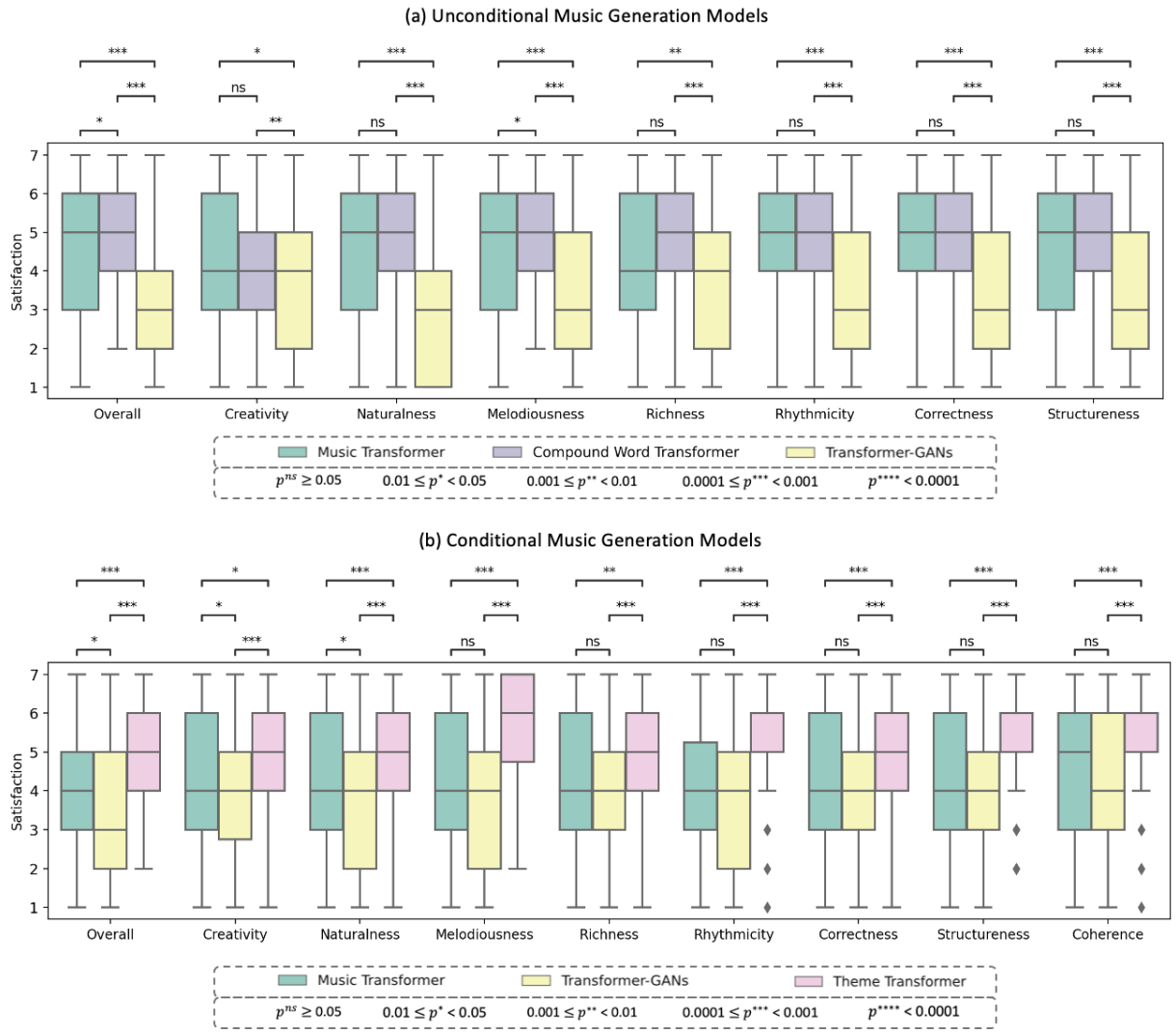

Figure 3: Satisfaction with Unconditional and Conditional Music Generation Models (1–7 Likert scale).

higher satisfaction in melodiousness than both Music Transformer (μ=4.713, SD=1.662, p=0.0117) and Transformer-GANs (μ=3.233, SD=1.772, p< 0.01). Furthermore, the qualitative analysis identified the metric receiving the most positive feedback as melodiousness (N=25), with abundant comments praising the melodiousness of Compound Word Transformer. For example, P_u 22 mentioned that *"The melody was super nice. I would listen to the song the whole day if I could."*, while P_u 2 stated that *"It felt very 'rich', with a good low-end presence and very nice, mellow melody"*. The results indicate that Compound Word Transformer achieves its goal of capturing and reflecting the co-occurrence relationships among different types of tokens to generate music in which listeners can feel a rich melody.

Considering the relative differences between musical dimensions creates rich songs, like human-composed music. Music Transformer recognizes that there are multiple dimensions in music and aims to reflect the relative differences among those dimensions in the music it generates. Therefore, the model captures the various relationships among different musical elements and uses them to generate rich songs. Our quantitative and qualitative results support the strength of this model in this regard. As Fig. 3 (a) shows, Music Transformer (μ=4.217, SD=1.675) received high score for **richness** in the unconditional music generation test. Tukey post-hoc analysis found that the model's richness score does not differ significantly from Compound Word Transformer (μ=4.690, SD=1.524, p=0.057), but is significantly higher than those of Transformer-GANs (μ=3.558, SD=1.729, p=0.004). In addition, several participants (N=14) expressed satisfaction with Music Transformer's its *richness*, saying, *"The songs have a wide range of sounds"*(P_u 33), and *"Different note heights added some richness to the sound"*(P_u 21). Furthermore, we found that this degree of richness enabled people to feel that the songs were natural. As Fig. 3 (b) shows, Music Transformer received high **naturalness** in both unconditional (μ=4.636, SD=1.698) and conditional (μ=4.3, SD=1.590) music generation test, although in unconditional test, the model's score did not differ significantly from those of Compound Word Transformer (μ=5.085, SD=1.398, p=0.0736).

Meanwhile, Music Transformer earned significantly higher *naturalness* than Transformer-GANs in both unconditional (μ=3.024, SD=1.798, p<0.001) and conditional (μ=3.7, SD=1.754, p=0.0114) test. Participants supported the results with comments such as, *"I wouldn't be able to find a difference between music played by an AI and music played by a real human."*(P_u 41) and even *"the songs can break into the commercial music industry"*(P_u 15).

Compact and diverse music development from one phrase to another enhances coherence. Transformer-GANs points out that the quality of music generated using an autoregressive model decreases when the model generates longer sequences. To address this issue, Transformer-GANs uses adversarial losses to maintain coherence in longer sequences. As a result, it can prevent songs from containing overly repetitive phrases or inconsistent development of their music. Our results support the strength of this model in this regard. As Fig 3 (b) shows, the model received the highest score for **coherence** (μ=4.192, SD=1.841) in conditional music generation test. Tukey post-hoc analysis found that the model received comparable satisfaction in *coherence* to Music Transformer (μ=4.517, SD=1.708, p=0.2703), as the two did not differ significantly. The participants' opinions were also in agreement with our qualitative analysis. For example, P_c 28 explained that he gave the highest overall score to Transformer-GANs because of *"the coherence of the music"*. Interestingly, several participants in the unconditional music generation test commented on the song's coherence even though we did not provide coherence as a metric in that test. These results indicate that Transformer-GANs achieves its aim of generating coherent music in longer sequences. As two participants noted, *"They had such a coherence"*(P_u 24) and, *"It was played quite orderly"*(P_u 27).

Theme-based conditioning magnifies structureness and melodiousness. Theme Transformer uses theme-based conditioning, which uses the music fragments that appear multiple times in the training set as input. To generate the next music sequence, it uses separate decoders for previously generated music sequences and the the provided conditions. As a result, Theme Transformer well reflects the theme to generate continuing sequences by making variations while preserving the theme. Our quantitative results support the model's the strength in this regard. As Fig. 3 (b) shows, Theme Transformer (μ=5.157, SD=1.498) received significantly higher scores in **structureness** than Music Transformer (μ=4.267, SD=1.662, p<0.001) and Transformer-GANs (μ=3.892, SD=1.580, p<0.001). Our qualitative analysis also found that theme-based music representation is more effective than prompt-based one in providing structureness. Several participants (N=9) cited the songs' structureness in comments, including *"a beautiful structure"*(P_c 19) and, *"It was nicely composed; structurally it felt absolutely professional"*(P_c 27).

Furthermore, we found that Theme Transformer's model characteristics also have a positive effect on **melodiousness**. As Fig. 3 (4) shows, the model received outstanding score for melodiousness (μ=5.442, SD=1.337), the highest score of all items regarding all models in both the unconditional and conditional test. Its score for melodiousness is significantly higher than those for Music Transformer (μ=4.358, SD=1.717, p<0.001) and Transformer-GANs (μ=3.867, SD=1.765, p<0.001). Several participants (N=16) commented on its melodiousness, including one who said, *"The songs had a nice melody. They actually had some repeating patterns, and it sounded pretty cohesive"*(P_c 28).

7 LESSONS LEARNED AND DISCUSSION

In this section, we elaborate on the lessons learned and discussion on AI-generated music and subjective evaluation criteria.

AI-generated music inspires creative activities. Creative activities include both musical and non-musical activities, such as writing, studying, and working. We found that the AI-generated songs inspire participants' musical creativity, including writing lyrics or dancing. For example, participants said that the songs created *"a desire to add a piece of lyrics to it"*(P_u 17) and *"to move my feet or mime along with it"*(P_u 14). Furthermore, we found that AI-generated music help people concentrate and focus when needed, such as when studying or reading (P_u 5). Another participant noted *"This kind of music is perfect when you are working on something that requires creativity and focus."* (P_u 41). Therefore, our results indicate that AI-generated music can boost several types of creative activities, including both musical and non-musical ones. In this context, exploring ways to enhance the creativity using various methods, including an interactive interface [19, 40, 41] could be a solution.

People want more controllability in AI music generation. We found that the context in which users listen to AI-generated music and their personal preferences affect their level of satisfaction with the songs. For example, P_u 18 thought, *"I would be more likely to listen to the song in another circumstance"*, while P_u 3 preferred *"songs with more instruments"* to those with a single instrument (e.g., piano). Therefore, providing options that enable users to select the listening circumstances or features of the music could improve human satisfaction with AI-generated music. Consequently, we can expect the models to include more controllability. In particular, we could enable users to freely select the tempo, pitch, genre, instruments, or length of music based on their circumstances and personal preferences. For example, users could select fast-tempo music before working out. Co-creation systems and interfaces for human-AI music composition (e.g., [40, 41]) can be beneficial because they enable users to customize the generated music based on their needs. Diverse options for enhancing users' control of the music composition process can better reflect users' contexts and preferences, thereby enriching their experiences.

AI-generated music enrich the experiences using other media. We found that people AI-generated music stimulate other sensory experiences (e.g., visual), as P_u 26 stated: *"I can well imagine them being used in the soundtrack of a game or movie though, with the right setting"*. Moreover, people express higher satisfaction with the songs that remind them of visual scenes, including *"The best score was for the one that made me feel like I was on a movie, like it could belong to the soundtrack of a period movie"* (P_c 31). The results show the potential of AI-generated music as background music for visual contents, including videos. Recently, several studies propose methods link visual and auditory experience: generating natural sound for in-the-wild videos [76], generating music from musical instrument playing video [20], or generating 3D dance movement from music [36]. Meanwhile, fewer studies propose methods to

Table 5: What is the most effective criteria?

	(a) Total		(b) Unconditional		(c) Conditional	
	User's Expectation	Actual Responses	User's Expectation	Actual Responses	User's Expectation	Actual Responses
1	Melodiousness (N=28)	Melodiousness (N=153)	Melodiousness (N=18)	Melodiousness (N=87)	Naturalness (N=13)	Melodiousness (N=66)
2	Naturalness (N=16)	Naturalness (N=98)	Rhythmicity (N=6)	Rhythmicity (N=47)	Melodiousness (N=10)	Naturalness (N=52)
3	Rhythmicity (N=13)	Rhythmicity (N=96)	Richness (N=6)	Naturalness (N=46)	Rhythmicity (N=7)	Rhythmicity (N=49)
4	Richness (N=10)	Creativity (N=70)	Creativity (N=6)	Creativity (N=43)	Richness (N=4)	Creativity (N=27)
5	Creativity (N=7)	Structureness (N=61)	Naturalness (N=3)	Richness (N=36)	Structureness (N=3)	Structureness (N=25)
6	Structureness (N=6)	Richness (N=51)	Structureness (N=3)	Structureness (N=36)	Coherence (N=2)	Correctness (N=20)
7	Correctness (N=1)	Correctness (N=46)	Correctness (N=1)	Correctness (N=26)	Creativity (N=1)	Coherence (N=19)
8	N/A	N/A	N/A	N/A	Correctness (N=0)	Richness (N=15)

generate background music for videos [11]. Through the our evaluation study, we can expect that AI-generated music can be used as soundtracks for a visual experience, such as a movie or a game, thereby enrich users' experience of other media.

Melodiousness is the most effective criterion, while Naturalness is still one of the most important ones. Table. 5 shows the most effective criteria based on the users' expectations and their actual responses. To figure out their expectations, we ask participants to select a criterion that they think is the most effective one to the overall satisfaction in our post survey questionnaire (Fig.2 (M6)). To discover the criterion that affects the most in reality, we count the number of mentions about the criterion. When counting the number, we include both positive and negative mentions because they all indicate that participants recognize that criterion in music. As shown in Table. 5, *melodiousness* wins all chart in user's expectation and actual response, including unconditional, conditional, and total. Moreover, *naturalness* and *rhythmicity* follows, as we can see in Table. 5(a). In section 4, we mentioned that we include the metrics to evaluate AI-generated music (i.e., *naturalness*, *correctness*, and *coherence*). Our results show that correctness and coherence do not have big impact on the users' satisfaction, as we expected. However, naturalness is one of the most effective criteria in both user's expectation and actual responses. Our qualitative results also support that users consider naturalness when evaluating AI-generated music. For example, P_u 41 gave a high score to a song because he *"wouldn't be able to find a difference between music played by an AI and music played by a real human"*. The results show that people tend to compare the AI-generated songs to the human-composed ones to estimate their level of satisfaction with AI-generated songs.

People value emotion, familiarity, and replayability when listening to AI-generated music. Participants suggested various elements of music other than the ones we proposed. The most frequently suggested were emotion (N=33), familiarity (N=5), and replayability (N=3). Our results showed that emotion is a significant aspect of music listening to music because, as one participant pointed out, *"There was no criterion to express how I felt while listening to the song"*(P_c 37). This indicates that a criterion regarding the emotional effect of the music is significant because it can estimate how one feels while listening to a song. Regarding *familiarity*, the comments indicated it can be defined as *"how familiar you are with this song"*(P_u 3) or *"whether you have heard something similar to the song"*(P_u 9). The results showed that participants enjoy songs more if they sounded familiar. The third suggested criterion was *replayability*, which can be defined as *"likelihood of listening to the songs again"*(P_u 29). While some studies have focused on objectifying the musical components (e.g., rhythm), our results indicate that people, especially novice users, pay more attention to their feelings. Therefore, developing subjective evaluation metrics related to emotion [18, 32, 35, 67–69] can better reflect the users' opinions.

Comprehensive evaluation metrics and detailed description help people understand AI-generated music. We found that our subjective metrics helped participants evaluate their satisfaction with various aspects of AI-generated music, because the metrics are *"vast"*(P_u 22) and *"complete"*(P_u 14). Also, the metrics also helped users understand the music, by making *"conscious and unconscious opinions clear"*(P_c 26). While prior studies [24] focus on the role of subjective metrics to estimate human satisfaction with the music, the results indicated that comprehensive evaluation metrics help people crystallize their feelings about and deepen their understanding of AI-generated music.

Moreover, many current subjective evaluation metrics for AI-generated music include professional terms (e.g., melody, pitch) [38, 39], but only a few of them have provided detailed explanations of each metric [1, 27]. We found that novice users depend on these descriptions to understand the meaning of each metric. For example, P_u 15 shared her personal experience as a beginner with little knowledge of music, saying that she *"had to move up and down to the definitions because I am not very musically informed"*. Users expressed satisfaction with the detailed description of each metric, saying *"The criteria seemed very well thought out and well explained"*(P_c 32) and *"The metrics are helpful even for somebody not educated in music"*(P_c 40). The results emphasized that to achieve a better evaluation, it is crucial to define each metric accurately and provide users with a detailed explanation of it.

8 CONCLUSION

Through analyzing 700 evaluations from 100 participants, we found that people have different perception towards various symbolic music generation models. The result shows that token representation and model characteristics make different satisfaction from people, in each of nine subjective evaluation metrics: (*overall*, *creativity*, *naturalness*, *melodiousness*, *richness*, *rhythmicity*, *correctness*, *structureness*, and *coherence*). We deepen our understanding of what people expect from AI-generated music, and how future AI-music generation system can meet the expectation.

ACKNOWLEDGEMENTS

This work was supported by the Korean National Research Foundation (NRF) grant (No. 2021R1A2C1004542) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No. 2020-0-01336–Artificial Intelligence Graduate School Program, UNIST), funded by the Korea government (MSIT).

REFERENCES

[1] Luca Angioloni, Tijn Borghuis, Lorenzo Brusci, and Paolo Frasconi. 2020. Conlon: A pseudo-song generator based on a new pianoroll, wasserstein autoencoders, and optimal interpolations. In *Proceedings of the 21th International Society for Music Information Retrieval Conference ISMIR MTL2020*. 876–883.

[2] Mira Balaban, Kermal Ebcioğlu, and Otto Laske. 1992. *Understanding music with AI: perspectives on music cognition*. MIT Press.

[3] Ethan Bresnick. 2019. Intensified Play: Cinematic study of TikTok mobile app. *University of Southern California* 4, 4 (2019), 1–12.

[4] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. 2017. Deep learning techniques for music generation–a survey. *arXiv preprint arXiv:1709.01620* (2017).

[5] Jean-Pierre Briot and François Pachet. 2020. Deep learning for music generation: challenges and directions. *Neural Computing and Applications* 32, 4 (2020), 981–993.

[6] Filippo Carnovalini and Antonio Rodà. 2020. Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence* 3 (2020), 14.

[7] Yu-Hua Chen, Yu-Hsiang Huang, Wen-Yi Hsiao, and Yi-Hsuan Yang. 2020. Automatic composition of guitar tabs by transformers and groove modeling. *arXiv preprint arXiv:2008.01431* (2020).

[8] Hang Chu, Raquel Urtasun, and Sanja Fidler. 2016. Song from PI: A musically plausible network for pop music generation. *arXiv preprint arXiv:1611.03477* (2016).

[9] Florian Colombo and Wulfram Gerstner. 2018. Bachprop: Learning to compose music in multiple styles. *arXiv preprint arXiv:1802.05162* (2018).

[10] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020).

[11] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. 2021. Video Background Music Generation with Controllable Music Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2037–2045.

[12] Chris Donahue, Huanru Henry Mao, Yiting Ethan Li, Garrison W Cottrell, and Julian McAuley. 2019. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. *arXiv preprint arXiv:1907.04868* (2019).

[13] Hao-Wen Dong, Ke Chen, Julian McAuley, and Taylor Berg-Kirkpatrick. 2020. MusPy: A toolkit for symbolic music generation. *arXiv preprint arXiv:2008.01951* (2020).

[14] 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.

[15] Hao-Wen Dong, Wen-Yi Hsiao, and Yi-Hsuan Yang. 2018. Pypianoroll: Open source Python package for handling multitrack pianoroll. In *Proc. Int. Soc. Music Information Retrieval Conf.*

[16] Hao-Wen Dong and Yi-Hsuan Yang. 2018. Convolutional generative adversarial networks with binary neurons for polyphonic music generation. *arXiv preprint arXiv:1804.09399* (2018).

[17] Jeff Ens and Philippe Pasquier. 2020. Mmm: Exploring conditional multi-track music generation with the transformer. *arXiv preprint arXiv:2008.06048* (2020).

[18] Lucas N Ferreira and Jim Whitehead. 2021. Learning to generate music with sentiment. *arXiv preprint arXiv:2103.06125* (2021).

[19] Emma Frid, Celso Gomes, and Zeyu Jin. 2020. Music creation by example. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[20] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. 2020. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*. Springer, 758–775.

[21] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 55–64.

[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] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1lYRjC9F7

[24] Carlos Hernandez-Olivan, Jorge Abadias Puyuelo, and Jose R Beltran. 2022. Subjective Evaluation of Deep Learning Models for Symbolic Music Composition. *arXiv preprint arXiv:2203.14641* (2022).

[25] Dorien Herremans, Ching-Hua Chuan, and Elaine Chew. 2017. A functional taxonomy of music generation systems. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 1–30.

[26] David Hesmondhalgh, Ellis Jones, and Andreas Rauh. 2019. SoundCloud and Bandcamp as alternative music platforms. *Social Media+ Society* 5, 4 (2019), 2056305119883429.

[27] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs. *arXiv preprint arXiv:2101.02402* (2021).

[28] Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck. 2019. Counterpoint by convolution. *arXiv preprint arXiv:1903.07227* (2019).

[29] 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).

[30] KC Huang, Q Jung, and J Lu. 2017. *Algorithmic music composition using recurrent neural networking*. Technical Report. Stanford University, Technical Report in CS221.

[31] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1180–1188.

[32] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. 2021. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. *arXiv preprint arXiv:2108.01374* (2021).

[33] Junyan Jiang, Gus G Xia, Dave B Carlton, Chris N Anderson, and Ryan H Miyakawa. 2020. Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 516–520.

[34] Jin Kim. 2012. The institutionalization of YouTube: From user-generated content to professionally generated content. *Media, culture & society* 34, 1 (2012), 53–67.

[35] Sang-gil Lee, Uiwon Hwang, Seonwoo Min, and Sungroh Yoon. 2017. Polyphonic music generation with sequence generative adversarial networks. *arXiv preprint arXiv:1710.11418* (2017).

[36] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.

[37] Feynman T Liang, Mark Gotham, Matthew Johnson, and Jamie Shotton. 2017. Automatic Stylistic Composition of Bach Chorales with Deep LSTM.. In *ISMIR*. 449–456.

[38] Xia Liang, Junmin Wu, and Jing Cao. 2019. MIDI-Sandwich2: RNN-based Hierarchical Multi-modal Fusion Generation VAE networks for multi-track symbolic music generation. *arXiv preprint arXiv:1909.03522* (2019).

[39] Hao-Min Liu and Yi-Hsuan Yang. 2018. Lead sheet generation and arrangement by conditional generative adversarial network. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 722–727.

[40] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[41] Ryan Louie, Jesse Engel, and Cheng-Zhi Anna Huang. 2022. Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation. In *27th International Conference on Intelligent User Interfaces*. 405–417.

[42] Elias Lousseief and Bob Sturm. 2019. Mahlernet: Unbounded orchestral music with neural networks. In *the Nordic Sound and Music Computing Conference 2019 and the Interactive Sonification Workshop 2019*. 57–63.

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

[44] Todd M. McComb. 2022. Medieval Music and Arts Foundation. http://www.medieval.org/emfaq/misc/homophony.html.

[45] Olof Mogren. 2016. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904* (2016).

[46] Aashiq Muhamed, Liang Li, Xingjian Shi, Suri Yaddanapudi, Wayne Chi, Dylan Jackson, Rahul Suresh, Zachary C Lipton, and Alexander J Smola. 2021. Symbolic music generation with Transformer-GANs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 408–417.

[47] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. 2020. This Time with Feeling: Learning Expressive Musical Performance. *Neural Comput. Appl.* 32, 4 (feb 2020), 955–967. https://doi.org/10.1007/s00521-018-3758-9

[48] Christine. Paynne. 2019. MuseNet. https://openai.com/blog/musenet/.

[49] Omar Peracha. 2019. Improving polyphonic music models with feature-rich encoding. *arXiv preprint arXiv:1911.11775* (2019).

[50] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. 2014. mir_eval: A transparent implementation of common MIR metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer.

[51] Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1198–1206.

[52] 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.

[53] 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).

[54] Andrew Shin, Leopold Crestel, Hiroharu Kato, Kuniaki Saito, Katsunori Ohnishi, Masataka Yamaguchi, Masahiro Nakawaki, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Melody generation for pop music via word representation of musical properties. *arXiv preprint arXiv:1710.11549* (2017).

[55] Aliaksandra Shutsko. 2020. User-generated short video content in social media. A case study of TikTok. In *International Conference on Human-Computer Interaction*. Springer, 108–125.

[56] Ian Simon and Sageev Oore. 2017. Performance RNN: Generating Music with Expressive Timing and Dynamics. https://magenta.tensorflow.org/performance-rnn.

[57] Ian Simon, Adam Roberts, Colin Raffel, Jesse Engel, Curtis Hawthorne, and Douglas Eck. 2018. Learning a latent space of multitrack measures. *arXiv preprint arXiv:1806.00195* (2018).

[58] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[59] Hao Hao Tan and Dorien Herremans. 2020. Music fadernets: Controllable music generation based on high-level features via low-level feature modelling. *arXiv preprint arXiv:2007.15474* (2020).

[60] TikTok. 2022. TikTok. https://www.tiktok.com/about.

[61] Andrea Valenti, Antonio Carta, and Davide Bacciu. 2020. Learning style-aware symbolic music representations by adversarial autoencoders. *arXiv preprint arXiv:2001.05494* (2020).

[62] José Van Dijck. 2009. Users like you? Theorizing agency in user-generated content. *Media, culture & society* 31, 1 (2009), 41–58.

[63] Sean Vasquez and Mike Lewis. 2019. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083* (2019).

[64] Federal Minimum Wage. 2022. Federal Minimum Wage. https://www.minimum-wage.org/federall.

[65] E. Waite, D. Eck, A. Roberts, and D. Abolafia. 2016. Project magenta: Generating long-term structure in songs and stories. https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn.

[66] Ziyu Wang, Yiyi Zhang, Yixiao Zhang, Junyan Jiang, Ruihan Yang, Junbo Zhao, and Gus Xia. 2020. Pianotree vae: Structured representation learning for polyphonic music. *arXiv preprint arXiv:2008.07118* (2020).

[67] Shih-Lun Wu and Yi-Hsuan Yang. 2020. The Jazz Transformer on the front line: Exploring the shortcomings of AI-composed music through quantitative measures. *arXiv preprint arXiv:2008.01307* (2020).

[68] Xianchao Wu, Chengyuan Wang, and Qinying Lei. 2020. Transformer-XL based music generation with multiple sequences of time-valued notes. *arXiv preprint arXiv:2007.07244* (2020).

[69] 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).

[70] Li-Chia Yang and Alexander Lerch. 2020. On the evaluation of generative models in music. *Neural Computing and Applications* 32, 9 (2020), 4773–4784.

[71] YouTube. 2022. YouTube. https://about.youtube/.

[72] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. Musicbert: Symbolic music understanding with large-scale pre-training. *arXiv preprint arXiv:2106.05630* (2021).

[73] Ning Zhang. 2020. Learning adversarial transformer for symbolic music generation. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[74] Jingwei Zhao and Gus Xia. 2021. AccoMontage: Accompaniment Arrangement via Phrase Selection and Style Transfer. *arXiv preprint arXiv:2108.11213* (2021).

[75] Yichao Zhou, Wei Chu, Sam Young, and Xin Chen. 2018. BandNet: A neural network-based, multi-instrument Beatles-style MIDI music composition machine. *arXiv preprint arXiv:1812.07126* (2018).

[76] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3550–3558.

[77] Hongyuan Zhu, Qi Liu, Nicholas Jing Yuan, Chuan Qin, Jiawei Li, Kun Zhang, Guang Zhou, Furu Wei, Yuanchun Xu, and Enhong Chen. 2018. Xiaoice band: A melody and arrangement generation framework for pop music. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2837–2846.