Affective and Pleasurable Design, Vol. 123, 2024, 82–93
https://doi.org/10.54941/ahfe1004679

AI-Driven Music Generation and Emotion Conversion

Xinwei Gao[1,2,3], Dengkai Chen[1,2], Zhiming Gou[1,2], Lin Ma[1,2], Ruisi Liu[1,2], Di Zhao[1,2], and Jaap Ham[3]

[1]Key Laboratory, Ministry of Industrial and Information Technology, Industrial Design Department, Northwestern Polytechnical University, Xi'an, 710072, China
[2]Shaanxi Engineering Laboratory for Industrial Design, Industrial Design Department, Northwestern Polytechnical University, Xi'an, 710072, China
[3]Research Group of Human Technology Interaction, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, 5600 MB, The Netherlands

ABSTRACT

With the integration of Generalized Adversarial Networks (GANs), Artificial Intelligence Generated Content (AIGC) overcomes algorithmic limitations, significantly enhancing generation quality and diversifying generation types. This advancement profoundly impacts AI music generation, fostering emotionally warm compositions capable of forging empathetic connections with audiences. AI interprets input prompts to generate music imbued with semantic emotions. This study aims to assess the accuracy of AI music generation in conveying semantic emotions, and its impact on empathetic audience connections. ninety audios were generated across three music-generated software (*Google musicLM*, *Stable Audio*, and *MusicGen)*, using four emotion prompts (Energetic, Distressed, Sluggish, and Peaceful) based on *the Dimensional Emotion Model*, and two generated forms (*text-to-music* and *music-to-music*). Emotional judgment experiment involving 26 subjects were conducted, comparing their valance and arousal judgments of the audios. Through Multi-way variance analysis, the AI-music-generated software had a significant main effect on the accuracy of conversion. Due to the diversity of generated forms of *MusicGen*, it has a lower accuracy of conversion compared to *Google musicLM* and *Stable Audio*. There was a significant interaction effect of generated forms and emotion prompts on the accuracy of conversion. The differences in accuracy between emotion prompts in the form of *text-to-music* were statistically significant, except for the differences between the accuracy of Distressed and Peaceful. Compared with the generated form of *text-to-music*, the form of *music-to-music* showed statistically significant emotional conversion ability for low arousal. The diversity of AI software input elements (i.e., text or music) may affect the effectiveness of emotional expression in music generation. The ability of different software to convey different emotions according to different prompts was unsteady in the form of *text-to-music.* This study advance computer music co-composition and improvisation abilities, facilitating AI music applications in fields such as medical rehabilitation, education, psychological healing, and virtual reality experiences.

Keywords: AI-music, Emotion conversion, Emotion model

INTRODUCTION

With the iterative progress of artificial intelligence (AI) technology, AI-generated content (AIGC) has attracted much attention in the field of computer science. Generative Artificial Intelligence (GAI) models can be categorized into five aspects: text generation, image generation, audio generation, video generation, and multimodal generation which accepts cross-modal instructions for random combinations. On November 30, 2022, ChatGPT, OpenAI's natural language processing model based on the Transformer framework, was launched, symbolizing the shift in the role of AI from an assistant tool for human creative work to an independent creative entity (Wu et al., 2023). As a result of technological innovation, AI music generation plays an increasingly important role in the future of the digital world.

Audio generation refers to the process of synthesizing corresponding sound wave forms based on input prompts (Dash and Agres, 2023). Broadly speaking, Artificial Intelligence Music (AI-Music) encompasses three categories: speech, music, and environmental sound. As an advanced stage in the development of AI music, the focus of AI music research and application is often "strongly technocratic", which tends to ignore the intrinsic connection between technological humanities and technological achievements themselves, such as aesthetic imagination, emotional cognition, and emotion generation (Xiao, 2022). In previous AI music research, scholars pay more attention to speech recognition, selection, and information algorithms. However, for the premise of AI music technology, whether AI music can make the audience empathize has gradually become a research trend (Wu et al., 2023).

Music possesses the capability to convey intense emotions. In studies of typical emotional expression in music, it has been found that audiences can recognize emotions such as happiness, sadness, anger, threat, and tenderness with a very high degree of accuracy (Vieillard et al., 2008). In previous music and emotion studies, researchers have primarily used clips of well-known Western classical music, such as the Baroque and Romantic periods (Laurier et al., n.d.). These clips are usually selected by researchers randomly and participants are likely to be already familiar with the music clips, which may cause emotional responses to be influenced by musical external affiliations. As a result, research on the effects of the selection, quality, and quantity of music clips as stimuli in studies is more limited (Kreutz et al., 2008). Although previous research has attempted to avoid these problems by using some synthetic music stimuli, these stimuli have difficulty in provoking emotional empathy with audiences due to the lack of complex features like expressiveness and tone that real music provides. However, with the breakthrough development of deep learning, big data, and cloud computing technologies, AI audio technology has entered a new stage of comprehensive development. AI music not only enhances expressive and tone effects as real music but also expresses emotions in a more comprehensive and in-depth way, thereby achieving a closer emotional connection with audiences.

Emotion is typically defined as a collection of psychological states that include subjective experiences, expressive behaviors, and surrounding physiological reactions. To create music that matches human emotional expression, AI music generation technology needs to learn models of human emotional expression (Laurier et al., n.d.). In the study of emotional theory, psychologists have proposed two typical models of human emotion: the Discrete emotion theory model and the Dimensional emotion model. Among them, the Dimensional emotion model has found widespread application in existing research on emotional models.

The Discrete Emotion Model

The Discrete emotion model is based on Tomkins' theory, Izard's model, and Ekman's model as the main theoretical frameworks (see Table 1). The theory suggests that all emotions are derivatives of a limited number of innate basic emotions, such as fear, anger, disgust, sadness, and joy (Ekman, 1992; Rubin and Talarico, 2009). In music and emotion research, basic emotion models are often modified to better describe the emotions expressed in music (Eerola and Vuoskoski, 2011). For example, emotions that are rarely expressed in music, such as disgust, are often replaced with more suitable emotional vocabulary, like gentle or serene. It has been argued that some basic emotions in *the Discrete emotion model* may seem insufficient to capture the richness of emotional effects in music (Zentner et al., 2008).

Table 1. Modern basic emotion theory.

Emotion Theory	Researchers	Emotional Categories
Basic emotion theory	Silvan Tomkins	Interest, enjoyment, surprise, fear, anger, pain, shame, contempt, disgust
Modern basic emotion theory	Paul Ekman	Anger, disgust, fear, joy, sadness, surprise
	Carroll Izard	Sadness, happiness, anger, fear, disgust, surprise, interest, shyness, self-condemnation, contempt
Psychoevolutionary theory	Robert Plutchik	joy vs sadness; anger vs fear; trust vs disgust; and surprise vs anticipation.

The Dimensional Emotion Model

The Dimensional models of emotion attempt to conceptualize human emotions by defining the degree of human emotions in two or three dimensions. In recent years, Russell's early Circumplex model - the Valence-Arousal (VA) two-dimensional emotion model, has been the most widely used nowadays (Gomez and Danuser, 2004). The two-dimensional circumplex model (Russell, 1980) suggests that emotions are distributed in a two-dimensional circular space, where all emotions can be represented by various degrees of valence and arousal (see Figure 1). The circumplex model is the most commonly used in testing stimuli related to emotion words, facial expressions, and emotional states. In 2021, Yik and Russell integrated a model of the core dimensions of mood and emotion based on the VA model and developed the

12-Point Affective Cycle (12-PAC) model of core emotions, which provided a clearer description of the core affective structure (Yik et al., 2011) (see Figure 2). In the field of music, Vieillard and others explored the emotional expression of music through similarity ratings, finding that these expressions could be mapped onto a two-dimensional model and well explained by two dimensions (Vieillard et al., 2008). In addition, the PAD three-dimensional model was developed by Albert Mehrabian and James A. Russell based on the two-dimensional emotion model, is one of the widely accepted three-dimensional psychological models for describing and measuring emotional states. As the current VA model aligns more with people's basic cognition of the objective world in real life and the emotion prompts in the VA two-dimensional model are more compatible with music's emotional expression, this research is based on the VA model for the analysis and exploration of AI music.

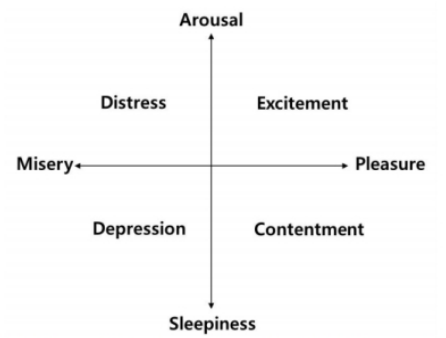

Fig. 1. Diagram of two-dimensional emotion model and four targeted emotional states.

Figure 1: The Valence-Arousal (VA) two-dimensional emotion model (adapted from Russell, 1980).

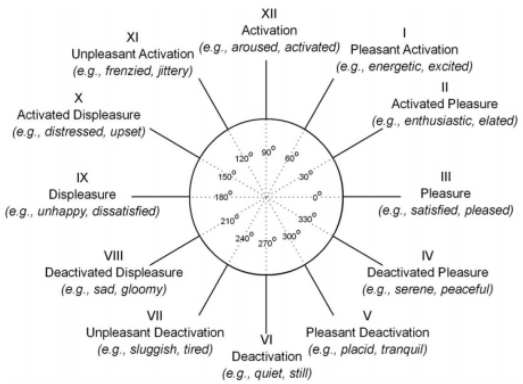

Figure 2: The 12-point affective cycle (12-PAC) model (adapted from Yik, 2011).

The research mentioned above indicates that music can be associated with and elicit emotional responses. AI music generation technology, as an

important branch of AIGC, its technology for speech synthesis and music cloning has been relatively mature, and the generated music presents effects and expressiveness that are extremely close to real music. With advances in artificial intelligence and machine learning, creating more emotional AI music will be a trend. To explore the emotional conveyance capacity of AI music generation, this paper will delve into the emotional aspects of AI music generation. Based on *the Dimensional emotion model* and experimental methods, the research will leverage three types of AI music generation software (*Google MusicLM, Stable Audio, MusicGen*) to investigate the emotion conveyance and empathetic effect of *text-to-music* and *music-to-music* in AI music generation.

EXPERIMENTAL METHODS AND PROCEDURES

Creators and Subjects

Twenty-six college students volunteered to participate in this experiment, including 6 females and 20 males (23.29±3.08). Four students were randomly selected as creators (two males and two females) in the experiment, with the remaining 22 students as subjects. All participants were not professionally trained in music appreciation but possessed some aesthetic qualities.

Experimental Materials

In this study, *the Dimensional emotion model* was chosen to explore the emotional delivery of AI music. Based on the 12-Point Affective Cycle (12-PAC) model, we categorized the four quadrants based on potency and arousal dimensions into three core emotion prompts according to different perspectives (see Table 2). Since the experimenters and participants were all Chinese graduate students, to consider the differences in cultural backgrounds, we conducted short interviews with all creators and subjects before the experiment to determine which emotion prompts elicited more intense emotional states. After interviews, the emotion prompts in the 12-PAC model with the angles of "60°, 150°, 240°, and 330°" were finally selected for further exploration, i.e., the four core emotion prompts of Energetic, Distressed, Sluggish, and Peaceful were used.

The creators chose the four core emotion prompts of Energetic (high valence-high arousal), Distressed (low valence-high arousal), Sluggish (low valence-low arousal), and Peaceful (high valence-low arousal) as the only emotion words in the prompts during the creation process and repeated them in the prompts to increase the weight of this keyword. The lexical category of the emotion prompts can be changed (e.g. energetic can be changed to energy). All creators were asked to limit their prompts to 20 words or less, with no adjectives other than the pre-defined emotion words. 15 prompts were generated by each emotion word, making a total of 60 prompts. For example: An energetic puppy running on the beach, the puppy is full of energy.

Table 2. A collection of emotion prompts from three equal angles.

Type	Angles	Core Emotion Vocabulary
1	0°, 90°, 180°, 270°	III(Satisfied, pleased), XII(aroused, activated), IX(unhappy, dissatisfied), VI(quiet, still).
2	30°, 120°, 210°, 300°	II(enthusiastic, elated), XI(frenzied, jittery), VIII(sad, gloomy), V(placid, tranquil).
	60°, 150°, 240°, 330°	I(energetic, excited), X(distressed, upset), VII(sluggish, tried), IV(serene, peaceful).
3	0°, 90°, 180°, 270°	III(Satisfied, pleased), XII(aroused, activated), IX(unhappy, dissatisfied), VI(quiet, still).

Due to the inconsistent duration among the three-generation software, we decided to splice or edit to standardize the audio length to 40s. Since *Stable audio* misinterpreted the prompts containing Distressed and output noise-dominant audio several times, we finally disregarded the effect of the music generation at low efficiency and high arousal in this software. In the form of *text to music*, 5 segments of corresponding AI music clips were generated by *Google musicLM*, *Stable audio*, and *MusicGen*, respectively, with a total of 55 segments of music. In addition, based on the generated audio, we randomly selected 10 pieces of each emotion type, making a total of 40 audios for secondary music creation. Since the *MusicGen* is the only software with the *music-to-music* function, we only use *MusicGen* to generate 10 segments of each of the corresponding emotion prompts, with a total of 40 audios, respectively. We chose 2–3 audios in each emotion category totaling 10 songs as pre-experimental material, and the remaining 85 pieces of music were used as formal experimental material (see Table 3).

Table 3. Statistics of experimental audio materials.

Forms		***Text-to-Music***			***Music-to Music***	**Total**
Software		Google musicLM	Stable audio	MusicGen	MusicGen	
Emotional words	Energetic	5	5	5	10	25
	Distressed	5	0	5	10	20
	Sluggish	5	5	5	10	25
	Peaceful	5	5	5	10	25
Total		20	15	20	40	95

Experimental Procedure

With the guidance of the four emotion types, the creators encoded their intentions and emotions into prompts and used AI music generation tools (*Google musicLM*, *Stable audio*, and *MusicGen*) to convey audio in both *text-to-music* and *music-to-music* forms. Finally, the subjects perceived their own emotions while listening (see Figure 3).

In the test, 85 mixed-emotion audios were played randomly and individually. The screen interface appears as a black circle to improve the subjects' experimental concentration. After completing the playing, the system will shift to the question interface automatically, and the subjects will use the digital keys 1–4 to make judgments about the emotion choices of the current audio (see Figure 4). The pre-experiment had the same process as the formal experiment to ensure the subjects' familiarity with and comprehension of the requirements and tempo of the test task. After listening, the audio will be evaluated by the subject at his/her speed and pace and there is no time limit. After evaluating the current audio, the program will present the next audio automatically. After evaluating every ten audios, subjects must take a two-minute break to avoid cognitive fatigue. At the end of all experiments, the subject will receive a small gift.

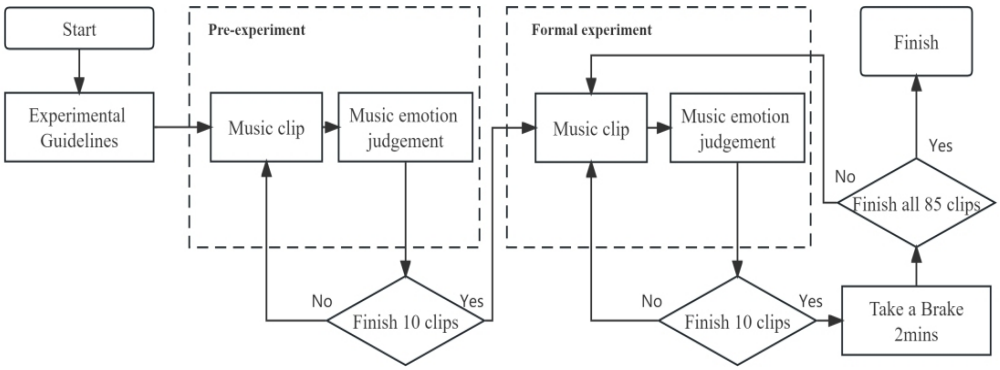

Figure 3: Experimental procedure.

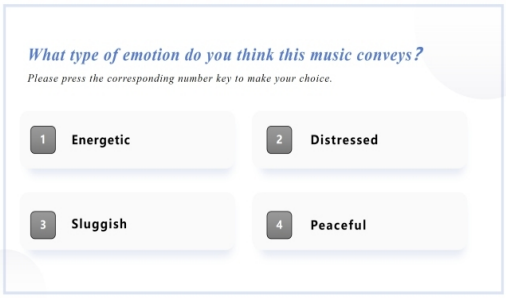

Figure 4: Interface of judgments about emotion choices.

The E-prime 3.0 program records subjects' choices of emotions automatically. By matching the emotion types judged by the subjects with the corresponding emotion prompts, successful matching is recorded as correct and failed matching is recorded as incorrect. The rate of emotion matching accuracy was obtained by analyzing the number of correct ones for all subjects in each generation form, generation software, and emotion prompts, thereby showing the effect of emotion transmission.

STATISTICS ANALYSIS

We summarized the emotion judgments into IBM SPSS Statistics 26 software to analyze the data. Kolmogorov-Smirnov (Kolmogorov-Smirnov, K-S) test and Q-Q plots were used to check the normality of the data distribution. Emotion judgment data were analyzed using a multi-way ANOVA method, which included generation forms (*text-to-music* & *music-to-music*), generation software (*Google musicLM*, *Stable audio*, *MusicGen*), and emotion prompts (Energetic, Distressed, Sluggish, Peaceful) and their interaction, using a two-side alpha level of 0.05 to determine statistical significance. Post-hoc tests were used when differences between independent variables were statistically significant. Since *Stable audio* could not regress the emotion prompt of Distressed, the interaction effects of generative model with generation forms and emotion prompts were not considered. Bonferroni method was used for a paried comparison of significant effects.

RESULT

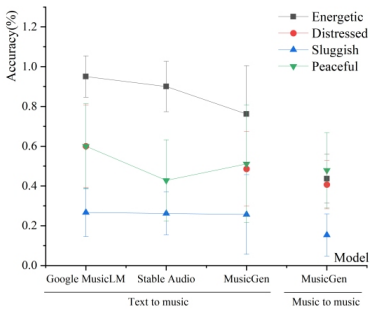

Figure 5: Descriptive statistics. Shows indicators of subjects' accuracy in judging the emotion of AI music under each music generation method and generation software. Since *MusicGen* software is the only software with Music to music function, so we show the data for only one software. Depressed was not considered for stable audio software.

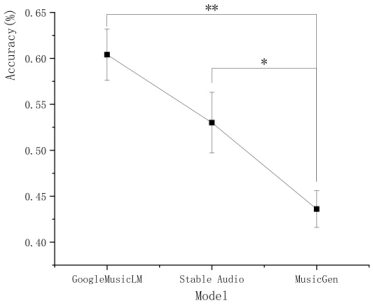

Figure 6: The AI-music-generated software has a significant main effect on the accuracy of conversion.

Energetic has a far higher accuracy than the other emotion prompts, with *Google musicLM* having the highest accuracy for this emotion prompt (0.95 ± 0.103), followed by *Stable audio* (0.90 ± 0.127). The accuracy of *MusicGen* in the form of *text-to-music* (0.762±0.242) was slightly higher than that in *music-to-music* (0.437±0.122), but the accuracy for this prompt in both forms was much lower than that of the other two types of software. *Google musicLM* was similar for Distressed (0.60±0.207) and Peaceful (0.60±0.213). *MusicGen*'s accuracy under *text-to-music* regarding Distressed (0.486±0.188) was similar to Peaceful (0.511±0.296), but the accuracy of Distressed (0.407±0.122) under *music-to-music* was lower than that of Peaceful (0.478±0.189). The processing of *Stable audio* (0.262±0.107) was the lowest for this word. Sluggish was significantly less accurate than the other emotion prompts, and the accuracy of this prompt under *music-to-music*(0.153±0.106) was significantly lower than that in *text-to-music*(0.261±0.120) (see Figure 5).

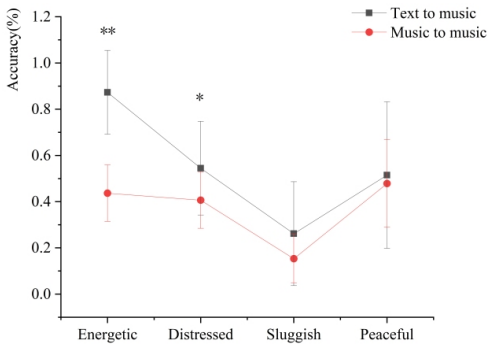

Figure 7: A Interaction effects between emotion prompts and generation forms.

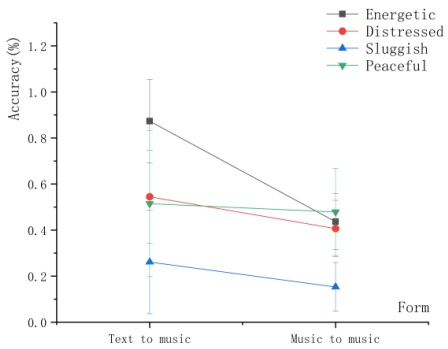

Figure 7: B Interaction effects between generation forms and emotion prompts.

The results of the analysis of variance (ANOVA) showed that the main effect of the factor "generating software" was significant (F = 12.579, P = 0.000, $\eta 2$=0.109). The difference between *Google musicLM* and *MusicGen* (0.436±0.020) was statistically highly significant (P<0.001), and

the difference between *Stable audio* and *MusicGen* was statistically significant (P<0.05). *Google musicLM* (0.604±0.028) generated significantly more than the other software generators, which was followed by *Stable audio* (0.530±0.033), and *MusicGen* (0.436±0.020) had the lowest generation effect(see Figure 6).

The interaction effect between emotion prompts and forms was significant (F = 2.666, P<0.05). The analysis of the simple effect of forms showed that the difference between the two generation forms was highly significant in the Energetic case (F = 45.366, P<0.001) and the effect of *text-to-music* (0.873±0.181) was significantly better than *music to music* (0.437±0.122). In the Distressed case, the difference in generation forms was significant (F = 3.950, P<0.05) and the effect of *text-to-music*(0.544±0.203) was better than *music-to-music* (0.407±0.122). As for Sluggish and Peaceful, all forms were not statistically significant (p>0.05), but the effect of *text-to-music* (0.545±0.203) was significantly better than that of *music-to-music* (0.407±0.122). The generative effect of Sluggish was the worst, and the generative effect of *text-to-music* (0.262±0.224) was slightly better than that of *music-to-music* (0.153±0.105), but still significantly lower than the generative effect of the other emotion prompts. The generative effect of Peaceful in *text-to-music* (0.515±0.307) was slightly better than that of *music-to-music* (0.479±0.189) (see Figure 7A).

The analysis of the simple effect of emotion prompts showed that the differences among emotion prompts were highly significant for *text-to-music* (F = 59.137, P< 0.001) and *music-to-music* (F = 6.960. P<0.001). The results of pairwise comparisons in the *text-to-music* showed that the differences between all emotion prompts were highly significant (P<0.001), except the differences between Distressed and Peaceful, which were not significant (P>0.05). Furthermore, the conveyance ability of Energetic (0.873±0.181) was extremely high, and the conveyance ability of Distressed (0.544±0.202) and Peaceful (0.515±0.307) was close to and higher than 50%, respectively. Compared Energetic (0.437±0.122) and Distressed (0.407±0.122) with Sluggish (0.153±0.105) respectively, the differences of both were significant (P<0.05) (see Figure 8).

DISCUSSION

This study investigates the ability of AI music generation to convey semantic emotions provided by the creators. In this study, 22 participants judged the emotion of AI music regarding four categories of emotional words, Energetic, Distressed, Sluggish, and Peaceful. The experiment verified that AI music has the ability to convey emotions. In the next step, we evaluated the accuracy of emotion judgment comprehensively, to compare the generation forms (*text-to-music* & *music-to-music*), generation software (*Google musicLM*, *Stable audio*, and *MusicGen*), and emotion prompts (Energetic, Distressed, Sluggish, Peaceful) for emotion conveyance effects.

Compared to the input of music, the input of text provides a better ability for emotion conveyance. In *text-to-music*, AI music has an extremely high ability in conveying emotion prompts with high valence and high arousal, an

average and similar ability in conveying high valence low arousal and high arousal low valence emotion prompts, and the weakest ability in transferring low valence low arousal emotion prompts. In *music-to-music*, the ability to convey high valence or high arousal mood types was weak.

The conveying ability of the three generative software is different and significantly differentiated, with *Google musicLM* having a great advantage in AI music generation, and we believe that possibly due to the diversity of input sources, the database training set leads to a weaker emotion conveyance in *MusicGen*.

In conclusion, AI music generation provides a more accurate conveyance of high arousal emotions (positive), but its ability to convey low arousal emotions (negative) needs to be improved. AI music is more specialized in processing text prompts than music prompts.

This study has limitations, and we plan to address these shortcomings in future research. The subjects were not surveyed using subjective questionnaires in this study, and the analysis of factors influencing emotions was solely based on the results of emotion judgments. In subsequent investigations, we aim to mitigate these limitations by incorporating subjective questionnaires to gather more comprehensive insights into the emotional responses of participants. Additionally, future studies will involve the selection of subjects from diverse age groups and academic backgrounds to further explore variations in the perception of AI music among individuals with different ages and disciplinary expertise.

ACKNOWLEDGMENT

Gratitude for the contributions of all the authors and the valuable insights from the reviewers.

REFERENCES

Dash, A., Agres, K. R., 2023. AI-Based Affective Music Generation Systems: A Review of Methods, and Challenges. https://doi.org/10.48550/ARXIV.2301.06890

Eerola, T., Vuoskoski, J. K., 2011. A comparison of the discrete and dimensional models of emotion in music. Psychology of Music 39, 18–49. https://doi.org/10.1177/0305735610362821

Ekman, P., 1992. Are There Basic Emotions. Psychol. Rev. 99, 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Gomez, P., Danuser, B., 2004. Affective and physiological responses to environmental noises and music. Int J Psychophysiol 53, 91–103. https://doi.org/10.1016/j.ijpsycho.2004.02.002

Kreutz, G., Ott, U., Teichmann, D., Osawa, P., Vaitl, D., 2008. Using music to induce emotions: Influences of musical preference and absorption. Psychology of Music 36, 101–126. https://doi.org/10.1177/0305735607082623

Laurier, C., Herrera, P., Fabra, U. P., n.d. Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines.

Rubin, D. C., Talarico, J. M., 2009. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. Memory 17, 802–808. https://doi.org/10.1080/09658210903130764

Russell, J. A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178. https://doi.org/10.1037/h0077714

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., Bouchard, B., 2008. Happy, sad, scary and peaceful musical excerpts for research on emotions. Cognition and Emotion 22, 720–752. https://doi.org/10.1080/02699930701503567

Wu, J., Gan, W., Chen, Z., Wan, S., Lin, H., 2023. AI-Generated Content (AIGC): A Survey.

Yik, M., Russell, J. A., Steiger, J. H., 2011. A 12-point circumplex structure of core affect. Emotion 11, 705–731. https://doi.org/10.1037/a0023980

Zentner, M., Grandjean, D., Scherer, K. R., 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. Emotion 8, 494–521. https://doi.org/10.1037/1528-3542.8.4.494

Xiao P, 2022. Embodiment, Imagination and Empathy: A Technophenomenological Study of Artificial Intelligence for Music Generation and Distribution. Modern Communication (Journal of Communication University of China) 44, 155–161. https://doi.org/10.19997/j.cnki.xdcb.2022.09.002